January 2015

Gaussian processes with built-in dimensionality reduction: Applications in high-dimensional uncertainty quantification

Rohit Kaushal Tripathy
Purdue University

Graduate School Form 30
Updated 1/15/2015

PURDUE UNIVERSITY
GRADUATE SCHOOL
Thesis/Dissertation Acceptance

This is to certify that the thesis/dissertation prepared

By Rohit Kaushal Tripathy

Entitled

Gaussian processes with built-in dimensionality reduction: Applications in high-dimensional uncertainty quantification

For the degree of Master of Science in Mechanical Engineering

Is approved by the final examining committee:

Ilias Bilionis
Chair

Alina Alexeenko

Marisol Koslowski

To the best of my knowledge and as understood by the student in the Thesis/Dissertation Agreement, Publication Delay, and Certification Disclaimer (Graduate School Form 32), this thesis/dissertation adheres to the provisions of Purdue University's "Policy of Integrity in Research" and the use of copyright material.

Approved by Major Professor(s): Ilias Bilionis

Approved by: Jay P. Gore 12/2/2015

Head of the Departmental Graduate Program Date

GAUSSIAN PROCESSES WITH BUILT-IN DIMENSIONALITY REDUCTION: APPLICATIONS IN HIGH-DIMENSIONAL UNCERTAINTY QUANTIFICATION

A Thesis

Submitted to the Faculty

of

Purdue University

by

Rohit K. Tripathy

In Partial Fulfillment of the

Requirements for the Degree

of

Master of Science in Mechanical Engineering

December 2015

Purdue University

West Lafayette, Indiana

To my parents, Bijay and Rashmi.

ACKNOWLEDGMENTS

I would like to begin by expressing my deepest gratitude to my advisor Prof. Ilias Bilionis. I am immensely grateful to have had the opportunity to work with him at the Predictive Science Lab here at Purdue for the past 18 months and I am immensely proud to continue working with him as a doctoral student. I would like to thank Prof. Marcial Gonzalez, whose simulation data has been crucial for the work laid down in this thesis, and Prof. Alejandro Strachan for his guidance during the weekly Uncertainty Quantification group meetings. I would also like to acknowledge the rest of the my thesis committee members, Prof. Alina Alexeenko and Prof. Marisol Koslowski, for their cooperation and support.

Finally, I would like to thank my labmates-Piyush and Nimish, current and former roommates- Nimish, Anurag and Anamitra, and all of my other friends for their support.

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

Figure Page

Figure Page

SYMBOLS

$\mathbb{R}$	set of reals
ϵ	random variable
δ	Dirac's delta
$\mathcal{D}$	data-set
$\mathbf{W}$	projection matrix
$V_d(\cdot)$	Stiefel Manifold
θ	vector of parameters
K	covariance matrix
$\mathcal{N}(\cdot,\cdot)$	gaussian distribution
$\mathbf{I}_N$	$N \times N$ Identity matrix
$\mathcal{L}$	loss function or likelihood
∇	gradient operator
τ	line search step-size

ABBREVIATIONS

UQ	uncertainty quantification
GP	Gaussian processes
AS	active subspace
MC	Monte Carlo
UP	uncertainty propagation
KLE	Karhunen Loeve expansion
PCA	principal component analysis
HDMR	high dimensional model representation
ANOVA	analysis of variance
PSL	partial least squares
BFGS	Broyden-Fletcher-Goldfarb-Shanno
PDE	partial differential equation
PDF	probability density function
SE	squared exponential
QoI	quantity of interest
MCMC	Markov chain Monte Carlo
MLE	maximum likelihood estimate
SVD	singular value decomposition
BIC	Bayesian Information Criterion

ABSTRACT

Tripathy, Rohit K. MSME, Purdue University, December 2015. Gaussian Processes with Built-In Dimensionality Reduction: Applications in High-Dimensional Uncertainty Quantification. Major Professor: Ilias Bilionis, School of Mechanical Engineering.

Uncertainty quantification (UQ) tasks, such as model calibration, uncertainty propagation, and optimization under uncertainty, typically require several thousand evaluations of the underlying computer codes. To cope with the cost of simulations, one replaces the real response surface with a cheap surrogate based, e.g., on polynomial chaos expansions, neural networks, support vector machines, or Gaussian processes (GP). However, the number of simulations required to learn a generic multivariate response grows exponentially as the input dimension increases. This curse of dimensionality can only be addressed, if the response exhibits some special structure that can be discovered and exploited. A wide range of physical responses exhibit a special structure known as an active subspace (AS). An AS is a linear manifold of the stochastic space characterized by maximal response variation. The idea is that one should first identify this low dimensional manifold, project the high-dimensional input onto it, and then link the projection to the output. If the dimensionality of the AS is low enough, then learning the link function is a much easier problem than the original problem of learning a high-dimensional function. The classic approach to discovering the AS requires gradient information, a fact that severely limits its applicability. Furthermore, and partly because of its reliance to gradients, it is not able to handle noisy observations. The latter is an essential trait if one wants to be able to propagate uncertainty through stochastic simulators, e.g., through molecular dynamics codes. In this work, we develop a probabilistic version of AS which is gradient-free and robust to observational noise. Our approach relies on a novel Gaussian process re-

gression with built-in dimensionality reduction. In particular, the AS is represented as an orthogonal projection matrix that serves as yet another covariance function hyper-parameter to be estimated from the data. To train the model, we design a two-step maximum likelihood optimization procedure that ensures the orthogonality of the projection matrix by exploiting recent results on the description of the tangent space of the Stiefel manifold, i.e., the manifold of orthogonal matrices. The additional benefit of our probabilistic formulation, is that it allows us to select the dimensionality of the AS via the Bayesian information criterion. We validate our approach by showing that it can discover the right AS in synthetic examples without gradient information using both noiseless and noisy observations. We demonstrate that our method is able to discover the same AS as the classical approach in a challenging one-hundred-dimensional problem involving an elliptic stochastic partial differential equation with random conductivity. Finally, we use our approach to study the effect of geometric and material uncertainties in shock propagation in a one dimensional granular system.

1. INTRODUCTION

Despite the indisputable successes of modern computational science and engineering, the increase in the predictive abilities of physics-based models has not been on a par with the advances in computer hardware. On one hand, we can now solve harder problems faster. On the other hand, however, the more realistic we make our models, the more parameters we have to worry about, in order to be able to describe boundary and initial conditions, material properties, geometric imperfections, constitutive laws, etc. Since it is typically impossible, or impractical, to accurately measure every single parameter of a complex computer code, we have to treat them as uncertain and model them using probability theory. Unfortunately, the field of uncertainty quantification (UQ) [1–4], which seeks to rigorously and objectively assess the impact of these uncertainties on model predictions, is not yet mature enough to deal with high-dimensional stochastic spaces.

The most straightforward UQ approaches are powered by Monte Carlo (MC) sampling [5, 6]. In fact, standard MC, as well as advanced variations, are routinely applied to the uncertainty propagation (UP) problem [7–9], model calibration [10,11], stochastic optimization [12–14], involving complex physical models. Despite the remarkable fact that MC methods convergence rate is independent of the number of stochastic dimensions, realistic problems typically require tens or hundreds of thousands of simulations. As stated by A. O'Hagan, this slow convergence is due to the fact that "Monte Carlo is fundamentally unsound" [15], in the sense that it fails to learn exploitable patterns from the collected data. Thus, MC is rarely ever useful in UQ tasks involving expensive computer codes.

To deal with expensive computer codes, one typically resorts to surrogates of the response surface. Specifically, one evaluates the computer code on a potentially adaptively selected, design of input points, uses the result to build a cheap-to-evaluate

version of the response surface, i.e., a surrogate. Then, he/she replaces all the occurrences of the true computer code in the UQ problem formulation with the constructed surrogate. The surrogate may be based on a generalized polynomial chaos expansion [16–20], radial basis functions [21, 22], relevance vector machines [23], adaptive sparse grid collocation [24], Gaussian Processes (GP) [23, 25–31] etc. For relatively low-dimensional stochastic inputs, all these methods outperform MC, that is that they need considerably fewer evaluations of the expensive computer code in order to yield satisfactorily convergent results.

In this work, we focus on Bayesian methods and, in particular, on GP regression [32]. The rationale behind this choice is due to the special ability of the Bayesian formalism to quantify the epistemic uncertainty induced by the limited number of simulations. In other words, it makes it possible to produce error bars for the results of the UQ analysis, see [23,28,30,31,33–37] and [38] for a recent review focusing on the uncertainty propagation problem. This epistemic uncertainty is the key to developing adaptive sampling methodologies, since it can be used to rigorously quantify the expected information content of future simulations. For example, see [39, 40] for adaptive sampling targeted to overall surrogate improvement, [41] and [42] for single- and multi-objective global optimization, respectively, and [28] for the uncertainty propagation problem.

Unfortunately, standard GP regression, as well as practically any generic UQ technique, is not able to deal with high stochastic dimensions. This is due to the fact that it relies on the Euclidean distance to define input-space correlations. Since the Euclidean distance becomes uninformative as the dimensionality of the input space increases [43], the number of simulations required to learn the response surface grows exponentially. This is known as the *curse of dimensionality*, a term coined by R. Bellman [44]. In other words, blindly attempting to learn generic high-dimensional functions is a futile task. Instead, research efforts are focused on methodologies that can identify and exploit some special structure of the response surface, which can be discovered from data.

The simplest way to address the curse of dimensionality is to use a variable reduction method, e.g., sensitivity analysis [45, 46] or automatic relevance determination [39, 47, 48]. Such methods rank the input features in order of their ability to influence the quantity of interest, and, then, eliminate the ones that are unimportant. Of course, variable reduction methods are effective only when the dimensionality of the input is reasonable (not very high-dimensional) and when the input variables are, more or less, uncorrelated. The common case of functional inputs, e.g., flow through porous media requires the specification of the permeability and the porosity as functions of space, cannot be teated directly with variable reduction methods. In such problems one has to start with a dimensionality reduction of the functional input. For example, if the input uncertainty is described via a Gaussian random field, dimensionality reduction can be achieved via a truncated Karhunen-Loève expansion (KLE) [49]. If the stochastic input model is to be built from data, one may use principal component analysis (PCA) [50], also known as empirical KLE, or even non-linear dimensionality reduction maps such as kernel PCA [51]. The end goal of dimensionality reduction techniques is the construction of a low dimensional set of uncorrelated features on which variable reduction methods, or alternative methods, may be applied. Note that even though the new features are lower dimensional than the original functional inputs, they are still high-dimensional for the purpose of learning the response surface.

A popular example of an exploitable feature of response surfaces that can be discovered from data is additivity. Additive response surfaces can be expressed as the sum of one-variable terms, two-variable terms, and so on, interpreted as interactions between combinations of input variables. Such representations are inspired from physics, e.g., the Coulomb potential of multiple charges, the Ising model of statistical mechanics. Naturally, this idea has been successfully applied to the problem of learning the energy of materials as a function of the atomic configuration. For example, in [14] the authors use this idea to learn the quantum mechanical energy of binary alloys on a fixed lattice by expressing it as the sum of interactions between clusters

of atoms, a response surface with thousands of input variables. The approach has also been widely used by the computational chemistry community, where it is known as high-dimensional model representation (HDMR) [52–55]. The UQ community has been embracing and extending HDMR [56, 57], sometimes referring to it by the name functional analysis of variance (ANOVA) [58, 59]. It is possible to model additive response surfaces with a GP by choosing a suitable covariance function. The first such effort can be traced to [60] and has been recently revisited by [61–65]. By exploiting the additive structure of response surfaces one can potentially deal with a few hundred to a few thousand input dimensions. This is valid, of course, only under the assumption that the response surface does have an additive structure with a sufficiently low number of important terms.

Another example of an exploitable response surface feature is active subspaces (AS) [66]. An AS is a low-dimensional linear manifold of the input space characterized by maximal response variation. It aims at discovering orthogonal directions in the input space over which the response varies maximally, ranking them in terms of importance, and keeping only the most significant ones. Mathematically, an AS is described by an orthogonal matrix that projects the original inputs to this low-dimensional manifold. The classic framework for discovering the AS was laid down by Constantine [67–70]. One builds a positive-definite matrix that depends upon the gradients of the response surface. The most important eigenvectors of this matrix form the aforementioned projection matrix. The dimensionality of the AS is identified by looking for sharp changes in the eigenvalue spectrum, and retaining only the eigenvectors corresponding to the highest eigenvalues. Once the AS is established, one proceeds by: i) Projecting all the inputs to the AS; ii) Learning the map between the projections and the quantity of interest. The latter is known as the *link function*. The framework has been successfully applied to a variety of engineering problems [71–75].

One of the major drawbacks of classic AS methodology is that it relies on gradient information. Even though, in principle, it is possible to compute the gradients either by deriving the adjoint equations [76] or by using automatic differentiation [77], in

many cases of interest this is not practical, since implementing any of these two approaches requires a significant amount of time for software development, validation and verification. This is an undesirable scenario when one deals with existing complex computer codes with decades of development history. The natural alternative of employing numerical differentiation is also not practical for high-dimensional input, especially when the underlying computer code is expensive to evaluate and/or when one has to perform the analysis using a restricted computational budget. The second major drawback of the classic AS methodology is its difficulty in dealing with relatively large observational noise, since that would require a unifying probabilistic framework. This drawback significantly limits the applicability of AS to important problems that include noise. For example, it cannot be used in conjunction with high-dimensional experimental data, or response surfaces that depend on stochastic models e.g., molecular dynamics.

The ideas of AS methodologies are reminiscent of the partial least squares (PSL) [78] regression scheme, albeit it is obvious that the two have been developed independently stemming from different applications. AS applications focus on computer experiments [67–70], while PSL has been extensively used to model real experiments with high-dimensional inputs/outputs in the field of chemometrics [79–81]. PSL not only projects the input to a lower dimensional space using an orthogonal projection matrix, but, if required, it can do the same to a high-dimensional output. It connects the reduced input to the reduced output using a linear link function. All model parameters are identified by minimizing the sum of square errors. PSL does not require gradient information and, thus, addresses the first drawback of AS. Furthermore, it also addresses, to a certain extent, the second drawback, namely the inability of AS to cope with observational noise, albeit only if the noise level is known a priori or fitted to the data using cross validation. As all non-Bayesian techniques, PSL may suffer from overfitting and from the inability to produce robust predictive error bars. Another disadvantage of PSL is the assumption that the link map is linear, a fact that severely limits its applicability to the study of realistic computer experiments.

The latter has been addressed by the locally weighted PSL [82], but at the expense of introducing an excessive amount of parameters.

In this work, we develop a probabilistic version of AS that addresses both its major drawbacks. That is, our framework is gradient-free (even though it can certainly make use of gradient information if this is available), and it can seamlessly work with noisy observations. It relies on a novel Gaussian process (GP) regression methodology with built-in dimensionality reduction. In particular, we treat the orthogonal projection matrix of AS as yet another hyper-parameter of the GP covariance function. That is, our proposed covariance function internally projects the high-dimensional inputs to the AS, and then models the similarity of the projected inputs. We determine all the hyper-parameters of our model, including the orthogonal projection matrix, by maximizing the likelihood of the observed data. To achieve this, we devise a two-step optimization algorithm guaranteed to converge to a local maximum of the likelihood. The algorithm iterates between the optimization of the projection matrix (keeping all other hyper-parameter fixed) and the optimization of all other hyper-parameters (keeping the projection matrix fixed), until a convergence criterion is met. To enforce the orthogonality constraint on the projection matrix, we exploit recent results on the description of the tangent space of the Stiefel manifold, i.e., the set of matrices with orthogonal columns. The optimization of the other hyper-parameters is carried out using BFGS [83]. The addendum of our probabilistic approach is that it allows us to select the dimensionality of the AS using the Bayesian information criterion (BIC) [84].

This paper is organized as follows. In Sec. 2.1, we briefly introduce GP regression, followed by a discussion of the classic, gradient-based, AS approach (Sec. 2.2) and the proposed gradient-free approach in (Sec. 2.3). Sec. 3.1 verifies our approach in a series of synthetic examples with known AS as well as it's robustness of our methodology to observational noise. In Sec. 3.2, we use a one-hundred-dimensional stochastic partial differential equation (PDE) to demonstrate that the proposed approach discovers the same AS as the classic approach - even without gradient information. In Sec. 3.3, we

use our approach to study the effect of geometric and material uncertainties in shock propagation through a one dimensional granular system. We present our conclusions in Ch. 4.

2. METHODOLOGY

Let $f : \mathbb{R}^D \to \mathbb{R}$ be a multivariate response surface with $D \gg 1$. Intuitively, $f(\cdot)$ accepts an *input*, $\mathbf{x} \in \mathbb{R}^D$, and responds with an *output* (or *quantity of interest* (QoI)), $f(\mathbf{x})$. We can measure $f(\mathbf{x})$ by querying an *information source*, which can be either a computer code or a physical experiment. Furthermore, we allow for noisy information sources. That is, we assume that instead of measuring $f(\mathbf{x})$ directly, we measure a noisy version of it $y = f(\mathbf{x}) + \epsilon$, where ϵ is a random variable. In physical experiments, measurement noise may rise from our inability to control all influential factors or from irreducible (aleatory) uncertainties. In computer simulations, measurement uncertainty may rise from quasi-random stochasticity, or chaotic behavior.

The ultimate goal of this work, is to efficiently propagate uncertainty through $f(\cdot)$. That is, given a probability density function (PDF) on the inputs:

$$\mathbf{x} \sim p(\mathbf{x}), \tag{2.1}$$

we would like to compute the statistics of the output. Statistics of interest are the *mean*

$$\mu_f = \int f(\mathbf{x}) p(\mathbf{x}) d\mathbf{x}, \tag{2.2}$$

the *variance*,

$$\sigma_f^2 = \int \left(f(\mathbf{x}) - \mu_f\right)^2 p(\mathbf{x}) d\mathbf{x}, \tag{2.3}$$

and the PDF of the output, which can be formally written as

$$f \sim p(f) = \int \delta\left(f - f(\mathbf{x})\right) p(\mathbf{x}) d\mathbf{x}, \tag{2.4}$$

where $\delta(\cdot)$ is Dirac's δ-function. We refer to this problem as the *uncertainty propagation* (UP) problem.

The UP problem is particularly hard when obtaining information about $f(\cdot)$ is expensive. In such cases, we are necessarily restricted to a limited set of observations. Specifically, assume that we have queried the information source at N input points,

$$\mathbf{X} = \left\{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(N)}\right\}, \tag{2.5}$$

and that we have measured

$$\mathbf{y} = \left\{y^{(1)}, \ldots, y^{(N)}\right\}. \tag{2.6}$$

We consider the following pragmatic interpretation of the UP problem: What is the best we can say about the statistics of the QoI, given the limited data in $\mathcal{D}$? The core idea behind our approach, and also behind most popular approaches in the current literature, is to replace the expensive response surface, $f(\cdot)$, with a cheap to evaluate surrogate learned from $\mathbf{X}$ and $\mathbf{y}$.

As discussed in Ch. 1, the fact that we are working in a high-dimensional regime, $D \gg 1$, causes insurmountable difficulties unless $f(\cdot)$ has some special structure that we can discover and exploit. In this work, we assume that the response surface has, or can be well approximated with the following form:

$$f(\mathbf{x}) \approx g(\mathbf{W}^T\mathbf{x}), \tag{2.7}$$

where the matrix $\mathbf{W} \in \mathbb{R}^{D\times d}$ projects the high-dimensional input space, $\mathbb{R}^D$, to the low-dimensional *active subspace*, $\mathbb{R}^d, d \ll D$, and $g : \mathbb{R}^d \to \mathbb{R}$ is a d-dimensional function known as the *link* function. Without loss of generality, we may assume that the columns of $\mathbf{W}$ are orthogonal. Mathematically, we write $\mathbf{W} \in V_d(\mathbb{R}^D)$, where $V_d(\mathbb{R}^D)$ is the set of $D \times d$ matrices with orthogonal columns,

$$V_d(\mathbb{R}^D) := \left\{\mathbf{A} \in \mathbb{R}^{D\times d} : \mathbf{A}^T\mathbf{A} = \mathbf{I}_d\right\}, \tag{2.8}$$

with $\mathbf{I}_d$ the $d \times d$ unit matrix, is known as the *Stiefel manifold*. Note that the representation of Equation (2.7) is arbitrary up to rotations and relabeling of the active subspace coordinate system. Intuitively, we expect that there is a d-dimensional subspace of $\mathbb{R}^D$ over which $f(\cdot)$ exhibits most of its variation. If d is indeed much smaller than D, then the learning problem is significantly simplified.

The goal of this paper is to construct a framework for the determination of the dimensionality of the active subspace d, the orthogonal projection matrix $\mathbf{W}$, and of the low dimensional map $g(\cdot)$ using only the observations $\{\mathbf{X}, \mathbf{y}\}$. Once these elements are identified, then one may use the constructed surrogate in any uncertainty quantification task, and, in particular, in the UP problem. We achieve our goal by following a probabilistic approach, in which $f(\cdot)$ is represented as a GP with $\mathbf{W}$ built into its covariance function and determined by maximizing the likelihood of the model.

2.1 Gaussian Process Regression

In this section we provide a brief, but complete, description of GP regression. Since, in later subsections, we use the concept in two different settings, here we attempt to be as generic as possible so that what we say is applicable to both. Towards this end, we consider the problem of learning an arbitrary response surface $h(\cdot)$ which takes inputs $\mathbf{q} \in \mathbb{R}^l$, assuming that we have made the, potentially noisy, observations:

$$\mathbf{t} = \left\{ t^{(1)}, \ldots, t^{(N)} \right\}, \tag{2.9}$$

at the input points:

$$\mathbf{Q} = \left\{ \mathbf{q}^{(1)}, \ldots, \mathbf{q}^{(N)} \right\}. \tag{2.10}$$

The philosophy behind GP regression is as follows. A GP defines a probability measure on a function space, i.e., a random field. This probability measure corresponds to our prior beliefs about the response surface. GP regression uses Bayes rule

to combine these prior beliefs with observations. The result of this process is a *posterior* GP which is simultaneously compatible with our beliefs and the data. We call this posterior GP a *Bayesian surrogate*. If a point-wise surrogate is required, one may use the median of the posterior GP. Predictive error bars, corresponding to the epistemic uncertainty induced by limited data, can be derived using the variance of the posterior GP. To materialize the GP regression program we need three ingredients: 1) A description of our prior state of knowledge about the response surface (Sec. 2.1.1); 2) A model of the measurement process (Sec. 2.1.2); and 3) A characterization of our posterior state of knowledge (Sec. 2.1.3). In Sec. 2.1.4 we discuss how the posterior of the model can be approximated via maximum likelihood.

2.1.1 Prior state of knowledge

Prior to seeing any data, we model our state of knowledge about $h(\cdot)$ by assigning to it a GP prior. We say that $h(\cdot)$ is a GP with mean function $m(\cdot;\boldsymbol{\theta})$ and covariance function $k(\cdot,\cdot;\boldsymbol{\theta})$, and write:

$$h(\cdot)|\boldsymbol{\theta} \sim \mathrm{GP}(h(\cdot)|m(\cdot;\boldsymbol{\theta}), k(\cdot,\cdot;\boldsymbol{\theta})). \tag{2.11}$$

The parameters of the mean and the covariance function, $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, are known as the *hyper-parameters* of the model.

Our prior beliefs about the response are encoded in our choice of the mean and covariance functions, as well as in the prior we pick for their hyper-parameters:

$$\boldsymbol{\theta} \sim p(\boldsymbol{\theta}). \tag{2.12}$$

The mean function is used to model any generic trends of the response surface, and it can have any functional form. If one does not have any knowledge about the trends of the response, then a reasonable choice is a zero mean function. The covariance function, also known as the covariance kernel, is the most important part of a GP.

Intuitively, it defines a nearness or similarity measure on the input space. That is, given two input points, their covariance models how close we expect the corresponding outputs to be. A valid covariance function must be positive semi-definite and symmetric. The most commonly used covariance function is the *square exponential* (SE):

$$k_{\mathrm{SE}}(\mathbf{q}, \mathbf{q}'; \boldsymbol{\theta}) = s^2 \exp\left\{ -\frac{1}{2} \sum_{i=1}^{l} \frac{(q_i - q_i')^2}{\ell_i^2} \right\}, \tag{2.13}$$

where $\boldsymbol{\theta} = \{s, \ell_1, \ldots, \ell_l\}$, with $s > 0$ being the signal strength and $\ell_i > 0$ the length scale of the i-th input. The SE covariance function corresponds to the a priori belief that the response surface is infinitely smooth. For more on covariance functions see Ch. 4 of Rasmussen [32].

Given an arbitrary set of inputs $\mathbf{Q}$, see Equation (2.10), Equation (2.11) induces, by definition, a Gaussian prior on corresponding response outputs:

$$\mathbf{h} = \left\{ h\left(\mathbf{q}^{(1)}\right), \ldots, h\left(\mathbf{q}^{(N)}\right) \right\}. \tag{2.14}$$

Specifically, $\mathbf{h}$ is a priori distributed according to:

$$\mathbf{h} | \mathbf{Q}, \boldsymbol{\theta} \sim \mathcal{N}\left(\mathbf{h} | \mathbf{m}, \mathbf{K}\right), \tag{2.15}$$

where $\mathcal{N}\left(\cdot | \boldsymbol{\mu}, \boldsymbol{\Sigma}\right)$ is the PDF of a multivariate Gaussian random variable with mean function $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, $\mathbf{m} := \mathbf{m}(\mathbf{Q}; \boldsymbol{\theta}) \in \mathbb{R}^N$ is the mean function evaluated at all points in $\mathbf{Q}$, i.e.,

$$\mathbf{m} = \mathbf{m}(\mathbf{Q}; \boldsymbol{\theta}) = \begin{pmatrix} m\left(\mathbf{q}^{(1)}; \boldsymbol{\theta}\right) \\ \vdots \\ m\left(\mathbf{q}^{(N)}; \boldsymbol{\theta}\right) \end{pmatrix}, \tag{2.16}$$

and $\mathbf{K} := \mathbf{K}(\mathbf{Q}, \mathbf{Q}; \boldsymbol{\theta}) \in \mathbb{R}^{N \times N}$ is the *covariance matrix*, a special case of the more general *cross-covariance matrix* $\mathbf{K}(\mathbf{Q}, \hat{\mathbf{Q}}; \boldsymbol{\theta}) \in \mathbb{R}^{N \times \hat{N}}$,

$$\mathbf{K}(\mathbf{Q}, \hat{\mathbf{Q}}; \boldsymbol{\theta}) = \begin{pmatrix} k\left(\mathbf{q}^{(1)}, \hat{\mathbf{q}}^{(1)}; \boldsymbol{\theta}\right) & \dots & k\left(\mathbf{q}^{(1)}, \hat{\mathbf{q}}^{(\hat{N})}; \boldsymbol{\theta}\right) \\ \vdots & \ddots & \vdots \\ k\left(\mathbf{q}^{(N)}, \hat{\mathbf{q}}^{(1)}; \boldsymbol{\theta}\right) & \dots & k\left(\mathbf{q}^{(N)}, \hat{\mathbf{q}}^{(\hat{N})}; \boldsymbol{\theta}\right) \end{pmatrix}, \tag{2.17}$$

defined between $\mathbf{Q}$, Equation (2.10), and an arbitrary set of $\hat{N}$ inputs $\hat{\mathbf{Q}} = \left\{\hat{\mathbf{q}}^{(1)}, \dots, \hat{\mathbf{q}}^{(\hat{N})}\right\}$.

2.1.2 Measurement process

The Bayesian formalism requires that we explicitly model the measurement process that gives rise to the observations $\mathbf{t}$ of Equation (2.9). The simplest such model is to assume that measurements are independent of each other, and that they are distributed normally about $h(\cdot)$ variance s_n^2. That is,

$$t^{(i)} | h\left(\mathbf{q}^{(i)}\right), s_n \sim \mathcal{N}\left(t^{(i)} \middle| h\left(\mathbf{q}^{(i)}\right), s_n^2\right). \tag{2.18}$$

Note that $s_n > 0$ is one more hyper-parameter to be determined from the data, and that we must also assign a prior to it:

$$s_n \sim p(s_n). \tag{2.19}$$

The assumptions in Equation (2.18) can be relaxed to allow for heteroscedastic (input dependent) noise [85, 86], but this is beyond the scope of this work. Using the independence assumption, we get:

$$\mathbf{t} | \mathbf{h}, s_n \sim \mathcal{N}\left(\mathbf{t} \middle| \mathbf{h}, s_n^2 \mathbf{I}_N\right). \tag{2.20}$$

Using the sum rule of probability theory and standard properties of Gaussian integrals, we can derive the *likelihood* of the observations given the inputs:

$$\mathbf{t}|\mathbf{Q},\boldsymbol{\theta},s_n \sim \mathcal{N}\left(\mathbf{t}\middle|\mathbf{m},\mathbf{K}+s_n^2\mathbf{I}_N\right). \tag{2.21}$$

2.1.3 Posterior state of knowledge

Using Bayes rule to combine the prior GP, Equation (2.11), with the likelihood, Equation (2.21), yields the *posterior* GP:

$$h(\cdot)|\mathbf{Q},\mathbf{t},\boldsymbol{\theta},s_n \sim \mathrm{GP}\left(h(\cdot)\middle|\tilde{m}(\cdot),\tilde{k}(\cdot,\cdot)\right), \tag{2.22}$$

where the *posterior* mean and covariance functions are

$$\tilde{m}(\mathbf{q}) := \tilde{m}(\mathbf{q};\boldsymbol{\theta}) = m(\mathbf{q};\boldsymbol{\theta}) + \mathbf{K}(\mathbf{q},\mathbf{Q};\boldsymbol{\theta})\left(\mathbf{K}+s_n^2\mathbf{I}_N\right)^{-1}(\mathbf{t}-\mathbf{m}), \tag{2.23}$$

and

$$\tilde{k}(\mathbf{q},\mathbf{q}') := \tilde{k}(\mathbf{q},\mathbf{q}';\boldsymbol{\theta},s_n) = k(\mathbf{q},\mathbf{q}';\boldsymbol{\theta}) - \mathbf{K}(\mathbf{q},\mathbf{Q};\boldsymbol{\theta})\left(\mathbf{K}+s_n^2\mathbf{I}_N\right)^{-1}\mathbf{K}(\mathbf{Q},\mathbf{q};\boldsymbol{\theta}), \tag{2.24}$$

respectively. The posterior of the hyper-parameters is obtained by combining Equation (2.12) and Equation (2.19) with Equation (2.20) using Bayes rule, i.e.,

$$\boldsymbol{\theta},s_n|\mathbf{Q},\mathbf{t} \sim p(\mathbf{t}|\mathbf{Q},\boldsymbol{\theta},s_n)p(\boldsymbol{\theta})p(s_n). \tag{2.25}$$

Equation (2.22) and Equation (2.25) fully quantify our state of knowledge about the response surface after seeing the data. However, in practice it is more convenient to work with the *predictive probability density* at a single input $\mathbf{q}$ conditional on the hyper-parameters $\boldsymbol{\theta}$ and s_n, namely:

$$h(\mathbf{q})|\mathbf{Q},\mathbf{t},\boldsymbol{\theta},s_n \sim \mathcal{N}\left(h(\mathbf{q})|\tilde{m}(\mathbf{q}),\tilde{\sigma}(\mathbf{q})\right), \tag{2.26}$$

where $\tilde{m}(\mathbf{q}) = \tilde{m}(\mathbf{q};\boldsymbol{\theta})$ is the predictive mean given in Equation (2.23), and

$$\tilde{\sigma}^2(\mathbf{q}) := \tilde{\sigma}^2(\mathbf{q},\mathbf{q}';\boldsymbol{\theta},s_n) = \tilde{k}(\mathbf{q},\mathbf{q}';\boldsymbol{\theta},s_n), \tag{2.27}$$

is the *predictive variance*. Note that the predictive mean can be used as a point-wise surrogate of the response surface, while the predictive variance can be used to derive point-wise predictive error bars.

2.1.4 Fitting the hyper-parameters

Ideally, one would like to characterize the posterior of the hyper-parameters, see Equation (2.25) using sampling techniques, e.g., a Markov chain Monte Carlo (MCMC) algorithm [87–89]. Here, we opt for a much simpler approach by approximating Equation (2.25) with a δ-Dirac function centered at the hyper-parameters that maximize the likelihood Equation (2.21). For issues of numerical stability, we prefer to work with the logarithm of the likelihood:

$$\mathcal{L}(\boldsymbol{\theta},s_n;\mathbf{Q},\mathbf{t}) := \log p(\mathbf{t}|\mathbf{Q},\boldsymbol{\theta},s_n). \tag{2.28}$$

and determine the hyper-parameters by solving the following optimization problem:

$$\boldsymbol{\theta}^*, s_n^* = \arg\max_{\boldsymbol{\theta},s_n} \mathcal{L}(\boldsymbol{\theta},s_n;\mathbf{Q},\mathbf{t}), \tag{2.29}$$

subject to any constraints imposed on the hyper-parameters(see Ch. 5 of [32]).

We refer to (Equation (2.21)) and express the likelihood as follows:

$$\log p(\mathbf{t}|\mathbf{Q},\boldsymbol{\theta},s_n^2) = -\frac{1}{2}(\mathbf{t}-\mathbf{m})^T\left(\mathbf{K}+s_n^2\mathbf{I}_N\right)^{-1}(\mathbf{t}-\mathbf{m}) - \frac{1}{2}\log|\mathbf{K}+s_n^2\mathbf{I}_N| - \frac{N}{2}\log 2\pi\ , \tag{2.30}$$

The derivatives of the likelihood with respect to any arbitrary parameter ϕ, where $\phi = s_n$ or $\phi = \boldsymbol{\theta}_i$ is obtained as follows:

$$\begin{aligned}\frac{\partial}{\partial \phi}\mathcal{L}(\boldsymbol{\theta}, s_n; \mathbf{Q}, \mathbf{t}) =& \frac{1}{2}\operatorname{tr}\Bigg(\Bigg(\Big(\mathbf{K} + s_n^2\mathbf{I}_N\Big)^{-1}\Big(\mathbf{t} - \mathbf{m}\Big) \\ & \Big(\Big(\mathbf{K} + s_n^2\mathbf{I}_N\Big)^{-1}\Big(\mathbf{t} - \mathbf{m}\Big)\Big)^T \\ & - \Big(\mathbf{K} + s_n^2\mathbf{I}_N\Big)^{-1}\Bigg)\frac{\partial \mathbf{K}}{\partial \phi}\Bigg)\end{aligned} \tag{2.31}$$

The point estimate of the hyper-parameters is known as the *maximum likelihood estimate* (MLE). The approach is justified if the prior is relatively flat and the likelihood is sharply picked. Unless otherwise stated, in this work we solve the optimization problem of Equation (2.29) via the BFGS optimization algorithm [83] increasing the chances of finding the global maximum by restarting the algorithm multiple times from random initial points.

It is completely possible that the log marginal likelihood, as a function of the hyperparameters, has multiple modes. It may, thus, exhibit multiple local optima. This does not seem to be a devastating problem. Different local optima correspond to different interpretations about the data. A practical solution would be to weight predictions from different interpretations according to their log likelihood and obtain an averaged prediction.

2.2 Gradient-Based Approach to Active Subspace Regression

In this section, we discuss the classic approach to discovering the active subspace using gradient information [67–70, 73–75, 90, 91]. Recall that we are dealing with a high-dimensional response surface, and that we would like to approximate it as in Equation (2.7). The classic approach does this in two steps. First, it identifies the projection matrix $\mathbf{W} \in V_d(\mathbb{R}^D)$ using gradient information (Sec. 2.2.1). Second, it

projects all inputs to the AS, and then uses GP regression to learn the map between the projected inputs and the output (Sec. 2.2.2).

Note that the classic approach is not able to deal with noisy measurments. Therefore, in this subsection, we assume that our measurements of $f(\mathbf{x})$ are exact. That is, we work under the assumption that each $y^{(i)}$ in Equation (2.6) is

$$y^{(i)} = f\left(\mathbf{x}^{(i)}\right), \tag{2.32}$$

for $i = 1, \ldots, N$. Also, since it requires gradient information, we assume that we have observations of the gradient of $f(\cdot)$ at each one of the input points, i.e., in addition to Equation (2.5) and Equation (2.6), we have access to:

$$\mathbf{G} = \left\{\mathbf{g}^{(1)}, \ldots, \mathbf{g}^{(N)}\right\}, \tag{2.33}$$

where

$$\mathbf{g}^{(i)} = \nabla f\left(\mathbf{x}^{(i)}\right) \in \mathbb{R}^D, \tag{2.34}$$

where $\nabla f(\cdot)$ is the gradient of $f(\cdot)$,

$$\nabla f(\cdot) = \left(\frac{\partial f(\cdot)}{\partial x_1}, \ldots, \frac{\partial f(\cdot)}{\partial x_D}\right). \tag{2.35}$$

2.2.1 Finding the active subspace using gradient information

Let $\rho(\mathbf{x})$ be a PDF on the input space, which can be different from the PDF of the UP problem given in Equation (2.1), and define the matrix

$$\mathbf{C} := \int (\nabla f(\mathbf{x}))(\nabla f(\mathbf{x}))^T \rho(\mathbf{x}) d\mathbf{x}. \tag{2.36}$$

Since $\mathbf{C}$ is symmetric positive definite, it admits the form

$$\mathbf{C} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^T, \tag{2.37}$$

where $\mathbf{\Lambda} = \text{diag}(\lambda_1, \cdots, \lambda_D)$ is a diagonal matrix containing the eigenvalues of $\mathbf{C}$ in decreasing order, $\lambda_1 \geq \ldots \geq \lambda_D \geq 0$, and $\mathbf{V} \in \mathbb{R}^{D \times D}$ is an orthonormal matrix whose columns correspond to the eigenvectors of $\mathbf{C}$. The classic approach suggests separating the d largest eigenvalues from the rest,

$$\mathbf{\Lambda} = \begin{bmatrix} \mathbf{\Lambda}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{\Lambda}_2 \end{bmatrix}, \quad \mathbf{V} = \begin{bmatrix} \mathbf{V}_1 & \mathbf{V}_2 \end{bmatrix},$$

(here $\mathbf{\Lambda}_1 = \text{diag}(\lambda_1, \ldots, \lambda_d), \mathbf{V}_1 = [\mathbf{v}_{11} \ldots \mathbf{v}_{1d}]$, and $\mathbf{\Lambda}_2, \mathbf{V}_2$ are defined analogously), and setting the projection matrix to

$$\mathbf{W} = \mathbf{V}_1^T. \tag{2.38}$$

Intuitively, $\mathbf{V}$ rotates the input space so that the directions associated with the largest eigenvalues correspond to directions of maximal function variability. See [67] for the theoretical justification.

It is impossible to evaluate Equation (2.36) exactly. Instead, the usual practice is to approximate the integral via Monte Carlo. That is, assuming that the observed inputs are drawn from $\rho(\mathbf{x})$, one approximates $\mathbf{C}$ using the observed gradients, see Equation (2.33), by:

$$\mathbf{C}_N = \frac{1}{N} \sum_{i=1}^{N} \mathbf{g}^{(i)} \left(\mathbf{g}^{(i)}\right)^T. \tag{2.39}$$

In practice, the eigenvalues and eigenvectors of $\mathbf{C}_N$ are found using the singular value decomposition (SVD) [92] of $\mathbf{C}_N$. The dimensionality d is determined by looking for sharp drops in the spectrum of $\mathbf{C}_N$.

2.2.2 Finding the map between the active subspace and the response

Using the classically found projection matrix, see Equation (2.38), we obtain the projected observed inputs $\mathbf{Z} \in \mathbb{R}^{N \times d}$:

$$\mathbf{Z} = \left\{ \mathbf{z}^{(1)}, \ldots, \mathbf{z}^{(N)} \right\}, \tag{2.40}$$

where

$$\mathbf{z}^{(i)} = \mathbf{W}^T \mathbf{x}^{(i)}. \tag{2.41}$$

The link function $g(\cdot)$ that connects the AS to the output, see Equation (2.7), is identified using GP regression, see Sec. 2.1, with response $h(\cdot) \equiv g(\cdot)$, input points $\mathbf{q} \equiv \mathbf{z}$, observed inputs $\mathbf{Q} \equiv \mathbf{Z}$, and observed outputs $\mathbf{t} \equiv \mathbf{y}$.

2.3 Gaussian Processes Regression with Built-In Dimensionality Reduction

As mentioned in Ch. 1 the classic approach to AS-based GP regression, see Sec. 2.2, suffers from two major drawbacks: 1) It relies on gradient information; and 2) It cannot deal seamlessly with measurement noise. In this section, we propose a probabilistic, unifying view of AS that is able to overcome these difficulties.

Our approach is based on novel covariance function on the high-dimensional input space:

$$k_{\mathrm{AS}} : \mathbb{R}^D \times \mathbb{R}^D \times V_d(\mathbb{R}^D) \times \boldsymbol{\Phi} \to \mathbb{R}, \tag{2.42}$$

with form:

$$k_{\mathrm{AS}}(\mathbf{x}, \mathbf{x}'; \mathbf{W}, \boldsymbol{\phi}) = k_d(\mathbf{W}^T \mathbf{x}, \mathbf{W}^T \mathbf{x}'; \boldsymbol{\phi}), \tag{2.43}$$

where $k_d : \mathbb{R}^d \times \mathbb{R}^d \times \boldsymbol{\phi} \to \mathbb{R}$ is a standard covariance function on the low-dimensional space parameterized by $\boldsymbol{\phi} \in \boldsymbol{\Phi}$. In words, the high-dimensional covariance function, Equation (2.43), first projects the inputs to the AS and, then, assesses the similarity of the projected inputs using the low-dimensional covariance function $k_d(\cdot, \cdot; \boldsymbol{\phi})$.

Note that he high-dimensional covariance function is parameterized by both the orthonormal projection matrix $\mathbf{W}$ and the hyper-parameters $\boldsymbol{\phi}$ of the low-dimensional covariance function.

To appreciate the unifying character of our approach note that the way to proceed is verbatim the generic GP regression approach of Sec. 2.1 with response $f(\cdot) \equiv h(\cdot)$, input points $\mathbf{q} \equiv \mathbf{x}$, observed inputs $\mathbf{Q} \equiv \mathbf{X}$, observed ouputs $\mathbf{t} \equiv \mathbf{y}$, covariance hyper-parameters $\boldsymbol{\theta} = \{\mathbf{W}, \boldsymbol{\phi}\}$ taking values in $\boldsymbol{\Theta} \equiv V_d(\mathbb{R}^D) \times \boldsymbol{\Phi}$, and covariance function $k(\cdot, \cdot; \boldsymbol{\theta}) \equiv k_{\mathrm{AS}}(\cdot, \cdot; \mathbf{W}, \boldsymbol{\phi})$. The only difficulty that we face, albeit non-trivial, that the likelihood maximization of Equation (2.29) must take into account the constraint that the projection matrix is orthonormal, $\mathbf{W} \in V_d(\mathbb{R}^D)$.

The rest of this subsection is concerned with the implementation of our paradigm. In Sec. 2.3.1 we present an iterative two-step likelihood maximization algorithm that is guaranteed to converge to a local maximum.

2.3.1 Iterative two-step likelihood maximization

Recall that a Gaussian Process is completely described by its hyperparameters $\boldsymbol{\theta}$. We expect data, when available in sufficient quantity, should be informative about these hyper-parameters. The process of training our Gaussian Process regression model is essentially an optimization problem which is posed as shown in Equation (2.29). In the present work, the low-dimensional kernel k_{AS} is characterized by an additional hyperparameter i.e. the projection matrix. We also established that the projection matrix is constrained to be orthogonal. We thus repose our optimization problem as follows:

$$\underset{\boldsymbol{\theta};W}{arg\,max} \mathcal{L}(\mathbf{W}, \boldsymbol{\theta}, s_n^2) = \underset{\boldsymbol{\theta};W}{arg\,max} \{\log p(\mathbf{t}|\mathbf{Q}, \mathbf{W}, \boldsymbol{\theta}, s_n^2)\} \ , \text{ s.t. } \mathbf{W}^T\mathbf{W} = \mathbf{I} \tag{2.44}$$

The minimization problem Equation (2.44) is high-dimensional which means that it would require a very large number of evaluations of the objective function in order to find the associated optima. The cost of evaluating the objective function is dictated

by the computation of an inverse which typically scales as $\mathcal{O}(N^3)$. We devise an iterative procedure that decomposes the high-dimensional optimization problem into a set of smaller problems. The result is of course suboptimal, but it can potentially yield good solutions.

We divide the optimization problem into a two-step method. We first randomly sample a projection matrix $\mathbf{W} \in V_d\left(\mathbb{R}^D\right)$ from a standard normal distribution and initialize the hyperparameters of the GP model. Keeping the hyper parameters fixed, we optimize the objective function $\mathcal{L}$, over the projection matrix $\mathbf{W}$. Note that this is a problem of orthogonality constrained minimization over a Stiefel manifold, and we talk about it in greater detail in the subsequent section. Once convergence is obtained, we turn our attention to the hyper parameters of the model. This time, we keep the optimized $\mathbf{W}$ matrix fixed, we now optimize $\mathcal{L}$ over the hyperparameters $\boldsymbol{\phi}$. Thus we optimize the marginal likelihood alternatively over the projection matrix $\mathbf{W}$ and the hyperparameters $\boldsymbol{\theta}$ and iterate until convergence. We summarize the process in Algorithm 1.

Algorithm 1 Algorithm to maximize the likelihood through a two-step iterative procedure

Require: Observed inputs $\mathbf{X}$, Observed outputs $\mathbf{y}$, maximum number of iterations maxitr, tolerance ϵ

Randomly initialize $\mathbf{W}$ by sampling each element independently from a standard normal.

Construct an orthonormal basis for $\mathbf{W}$ using the Singular Value Decomposition(SVD)

Initialize $\boldsymbol{\theta}$ and $s_n$by sampling from the prior(if it exists) else sample from a standard normal distribution.

Set $\mathbf{W}_0^* \leftarrow \mathbf{W}$

Set $\boldsymbol{\theta}_0^* \leftarrow \boldsymbol{\theta}$

Set $s_{n,0}^* \leftarrow s_n$

Set $\mathcal{L}_0 = \mathcal{L}(\mathbf{W}, \boldsymbol{\theta}, s_n)$

$i \leftarrow 1$

while $i <=$ maxitr **do**

  $\mathbf{W}_i^* \leftarrow \underset{W}{arg\,min} \mathcal{L}(\mathbf{W}_{i-1}^*, \boldsymbol{\theta}_{i-1}^*, s_{n,i-1}^*)$ using Alg. 2

  $\boldsymbol{\theta}_i^*, s_{n,\,i}^* \leftarrow \underset{\boldsymbol{\theta}}{arg\,min} \mathcal{L}(\mathbf{W}_i^*, \boldsymbol{\theta}_{i-1}^*, s_{n,i-1}^*)$ using BFGS algorithm [83]

  Set $\mathcal{L}_i \leftarrow \mathcal{L}(\mathbf{W}_i^*, \boldsymbol{\theta}_i^*, s_{n,i}^*)$

  if $\mathcal{L}_i - \mathcal{L}_{i-1} < \epsilon \mathcal{L}_{i-1}$ **then**

    break

  end if

  Set $i \leftarrow i + 1$

end while

return $\mathbf{W}_i^*, \boldsymbol{\theta}_i^*, s_{n,i}^*$

2.3.2 Maximizing the likelihood with respect to the projection matrix

The Stiefel manifold is a constrained submanifold in the $\mathbb{R}^{n\times p}$ space and is defined as the feasible set in Equation (2.8). The problem of finding a minimizer for a function defined over a Stiefel manifoldis formally expressed as follows:

$$\mathbf{W}^* = \arg \max_{\mathbf{W}\in V_d(\mathbb{R}^D)} \mathcal{L}(\mathbf{W}), \tag{2.45}$$

where,

$$\mathcal{L}(\mathbf{W}) = \mathcal{L}(\mathbf{W}, \boldsymbol{\theta}, s_n; \mathbf{X}, \mathbf{y}) \tag{2.46}$$

We note that the marginal likelihood is a function of the hyperparameters of the GP model as well the projection matrix. Initial estimates of $\mathbf{W}$ are obtained by orthogonalizing random matrix with normally distributed elements. The optimization procedure is based on the method proposed by [93]. Optimization over a constrained manifold is challenging oweing to non-convexity and difficulty in constraint preservation. The line-search algorithm employed is based on the Cayley transform, a Crank-Nicholson like update and is an efficient procedure to bypass these problems. In this section we briefly discuss the algorithm.

Given a feasible point $\mathbf{W}$, line search step size τ and the gradient $\mathbf{G} := \nabla_{\mathbf{W}}\mathcal{L}(\mathbf{W})$, we define a skew symmetric matrix $\mathbf{A}$ as follows:

$$\mathbf{A} := \mathbf{G}\mathbf{W}^T - \mathbf{W}\mathbf{G}^T \ . \tag{2.47}$$

The matrix $\mathbf{A}$ is obtained by taking the partial derivative, w.r.t. $\mathbf{W}$, of the objective function corresponding to the given constrained optimization problem (Equation (2.45)). The matrix $\mathbf{AW}$ represents the gradient of the Lagrangian function and the direction of the steepest descent. The natural idea for the next step would be to update $\mathbf{Y} := \mathbf{W} - \tau\mathbf{A}\mathbf{W}$. This scheme, however, does not guarantee preservation of

orthogonality. Instead, we update the trail point according to the following relation:

$$\mathbf{Y}(\tau) = \mathbf{P}\mathbf{W} \ , \tag{2.48}$$

where,

$$\mathbf{P} := (\mathbf{I} + \frac{\tau}{2}\mathbf{A})^{-1}(\mathbf{I} + \frac{\tau}{2}\mathbf{A}) \ . \tag{2.49}$$

This Crank-Nicholson like update scheme is known as a Cayley Transformation. It can be easily verified that the matrix $\mathbf{P}$ and therefore, the updated trail point $\mathbf{Y}$ are orthogonal matrices. Thus, the update preserves orthogonality.

In order to obtain gradients of the objective function with respect to the projection matrix we use Equation (2.31). To complete the process we need the gradients of the covariance matrix with respect to the projection matrix. We obtain these as follows:

$$\frac{\partial k_{nm}}{\partial w_{ij}} = -2k_{nm}\left(\mathbf{w}_i^T\mathbf{x}^{(n)} - \mathbf{w}_i^T\mathbf{x}^{(m)}\right)\left(x_j^{(n)} - x_j^{(m)}\right), \tag{2.50}$$

where k_{nm} and w_{ij} are the elements of $\mathbf{K}$ and $\mathbf{W}$, respectively and $\mathbf{w}_i$ is the i^{th} column of the matrix $\mathbf{W}$.

We note that the updated projection matrix estimate for each step of the iterative scheme is a function of the step size τ i.e. $\mathbf{W} = \mathbf{W}(\tau)$ and by extension the objective function is also a function of the step size τ.

$$\mathcal{L}(\mathbf{W}) = \mathcal{L}(\mathbf{W}, \tau, \boldsymbol{\theta}, s_n; \mathbf{X}, \mathbf{y}) \tag{2.51}$$

The objective function $\mathcal{L}$ is evaluated at the current trial point and the step size is updated to that value which minimizes $\mathcal{L}$. We use the Brent Algorithm [94] for this step. Our code provides the flexibility of using any of 3 different algorithms - Brent, Golden and Bounded. Indeed it should be noted that any algorithm that minimizes a scalar valued function may be used. Once the new step size is obtained the next trial point is obtained. This iterative process is carried out until convergence is obtained

based on a user specified tolerance. Algorithm 2 shows the pseudo code to minimize the objective function over the Stiefel Manifold.

Algorithm 2 Optimization over Stiefel Manifold

Require: Objective Function $\mathcal{L}(\mathbf{W})$ and its gradient $\mathbf{G}(\mathbf{W})$, Step Size τ, Maximum number of iterations $maxitr$, Stiefel Optimization tolerance ϵ

Randomly sample an initial trial point; $\mathbf{W}_0 \leftarrow \text{randn}(D, d)$

Orthogonalize $\mathbf{W}_0$ using the Singular Value Decomposition(SVD); $\mathbf{W} \leftarrow \text{orth}(\mathbf{W})$

$i \leftarrow 1$

while $i <= maxitr$ **do**

  Set $\mathbf{W}_i \leftarrow \mathbf{W}_{i-1}$

  Evaluate $\mathcal{L}_i, \mathbf{G}_i$ at $\mathbf{W}_i$

  Evaluate $\mathbf{A}$ according to Equation (2.47)

  Set $\tau^* \leftarrow \underset{\tau}{arg\,max}\mathcal{L}(\mathbf{W}, \tau)$ using the Brent algorithm [94]

  Set $\mathbf{W}^* \leftarrow \mathbf{Y}(\tau^*)$

  Set $\mathcal{L}^* \leftarrow \mathcal{L}(\mathbf{W}^*)$

  if $\mathcal{L}^* - \mathcal{L}_i < \epsilon\mathcal{L}_i$ **then**

    break

  end if

  $i \leftarrow i + 1$

end while

return $\mathbf{W}^*$

2.3.3 Identification of active subspace dimension

Algorithm 3 Identification of active subspace dimension

Require: Observed inputs $\mathbf{X}$, Observed outputs $\mathbf{y}$, marginal likelihood function $\mathcal{L}$, tolerance δ

1: $\text{BIC}_0 \leftarrow 0$
2: Set $d \leftarrow 1$
3: **repeat**
4: Train model using Alg. (1) and Alg. (2) with dimension of the active subspace set to d and obtain $\mathbf{W}^*, \boldsymbol{\theta}^*$ and s_n^*.
5: Obtain k from Equation (2.53)
6: **until** $\text{BIC}_d - \text{BIC}_{d-1} <= \delta\, \text{BIC}_{d-1}$
7: **return** $\mathbf{W}^*$

Statistical models are judged based on their out-of-sample predictive accuracy. Typically, when choosing between a finite set of models, we assign scores to each model based on a maximum likelihood estimate and penalize the score to compensate for model complexity in order to avoid over-fitting. A review on various scores used for assessing the predictive capabilities of statistical models can be found in Gelman [95]. In the present work, we need to select the lower dimensional mapping that represents the true active subspace of a given high dimensional data-set. We use the Bayesian Information Criterion or BIC score to compare models representing a range of lower dimensional representation of our data. The BIC score is defined as follows:

$$\text{BIC} = \log \mathcal{L} - k \log N \tag{2.52}$$

where, the first term $\log \mathcal{L}$ represents the logarithm of the marginal likelihood of the model, k is the number of free parameters in the models i.e., the number of hyperparameters, and n is the number of samples in the data-set. As is evident from Equation (2.52) the second term is a penalty term and it penalizes models that are

more complex. For any given low dimensional representation of the data-set we can compute k as follows:

$$k = d(D + 1) + 2 \tag{2.53}$$

In order to select the correct active dimension, we look at the following criterion:

$$\frac{\text{BIC}_{d+1} - \text{BIC}_d}{\text{BIC}_d} < \delta \tag{2.54}$$

The inequality 2.54 represents the relative increase in the BIC score as we increase the number of active dimensions in our model from d to d+1. The RHS δ is a threshold limit chosen arbitrarily by the user. When this criterion is met, we select d as the accurate dimensionality of the active subspace and terminate the model selection procedure.

3. EXAMPLES

We have implemented both the classic approach, Sec. 2.2, and the novel gradient-free approach, Sec. 2.3, in Python. Our code extends the celebrated GPy module [96] and is publicly available at https://bitbucket.org/rohitkt10/active_subspace_work/. All the numerical results we present here can be replicated by following the instructions on the aforementioned website.

Sec. 3.1 uses a series of synthetic examples (known projection matrix and known non-linear link function) to verify that the proposed approach, Sec. 2.3, finds the same AS the classic approach, Sec. 2.2. Our goal is to address the first identified drawback of classic AS, namely the reliance on gradient information. Furthermore, this section validates our claim that the proposed methodology is robust to measurement noise. In Sec. 3.2, we apply our technique to a standard UQ benchmark with one hundred input dimensions, a stochastic elliptic partial differential equation with random conductivity. The results are again compared to the classic AS, thereby verifying the agreement between the two in a more challenging, truly high-dimensional setting. We conclude this section with an exhaustive uncertainty analysis of a 1D granular crystal with geometric and material imperfections, see Sec. 3.3. The latter is not amenable to the classic AS approach due to lack of gradient information. Note that, to the best of our knowledge, this is the first time an uncertainty analysis of this scale has been performed to a granular crystal system.

Note that the output variables in all the plots throughout this section have been scaled using the following formula:

$$y_{\text{scaled}} = \frac{y - \mu_y}{\sigma_y}. \tag{3.1}$$

Here μ_y and σ_y are the mean and standard deviation of the output respectively.

3.1 Synthetic Response Surface with Known Structure

Let $f : \mathbb{R}^D \to \mathbb{R}$ be a response surface of the form:

$$f(\mathbf{x}) = g(\mathbf{W}^T\mathbf{x}), \tag{3.2}$$

with $\mathbf{W} \in V_d(\mathbb{R}^D)$, and quadratic link function $g : \mathbb{R}^d \to \mathbb{R}$,

$$g(\mathbf{z}) = \alpha + \boldsymbol{\beta}^T\mathbf{z} + \mathbf{z}^T\boldsymbol{\Gamma}\mathbf{z}, \tag{3.3}$$

with $\alpha \in \mathbb{R}, \boldsymbol{\beta} \in \mathbb{R}^d$ and $\boldsymbol{\Gamma} \in \mathbb{R}^{d\times d}$. The gradient of Equation (3.2) with respect to $\mathbf{x}$ is:

$$\nabla f(\mathbf{x}) = \left(\boldsymbol{\beta} + 2\mathbf{x}^T\mathbf{W}\boldsymbol{\Gamma}\right)\mathbf{W}^T. \tag{3.4}$$

In all the cases considered in this subsection, the number of input dimensions is ten, $D = 10$. The parameters $\mathbf{W}, \boldsymbol{\alpha}, \boldsymbol{\beta}$ and $\boldsymbol{\Gamma}$ were randomly generated. Reproducibility is ensured by fixing a random seed. Due to lack of space, we only give the values of these parameters when the dimension of the active subspace, d, is lower than or equal to two. For all other cases, we refer the reader to the accompanying website of this paper. Given a frozen set of parameters, we query the response $f(\cdot)$ at N normally distributed input points and contaminate the measurements with synthetic zero mean Gaussian noise with standard deviation $s_n > 0$. This results in a collection of inputs, $\mathbf{X}$ as in Equation (2.5), and outputs, $\mathbf{y}$ as in Equation (2.6). When needed, we also collect gradient data, $\mathbf{G}$ as in Equation (2.33), but we do not contaminate them with noise.

In Sec. 3.1.1 and Sec. 3.1.2, we verify that the gradient-free approach discovers the underlying 1D and 2D AS structure, respectively. Sec. 3.1.3 demonstrates the efficacy of the BIC as automatic method for the determination of the dimensionality of the AS. Finally, in Sec. 3.1.4 we study the robustness of the gradient-free approach to measurment noise.

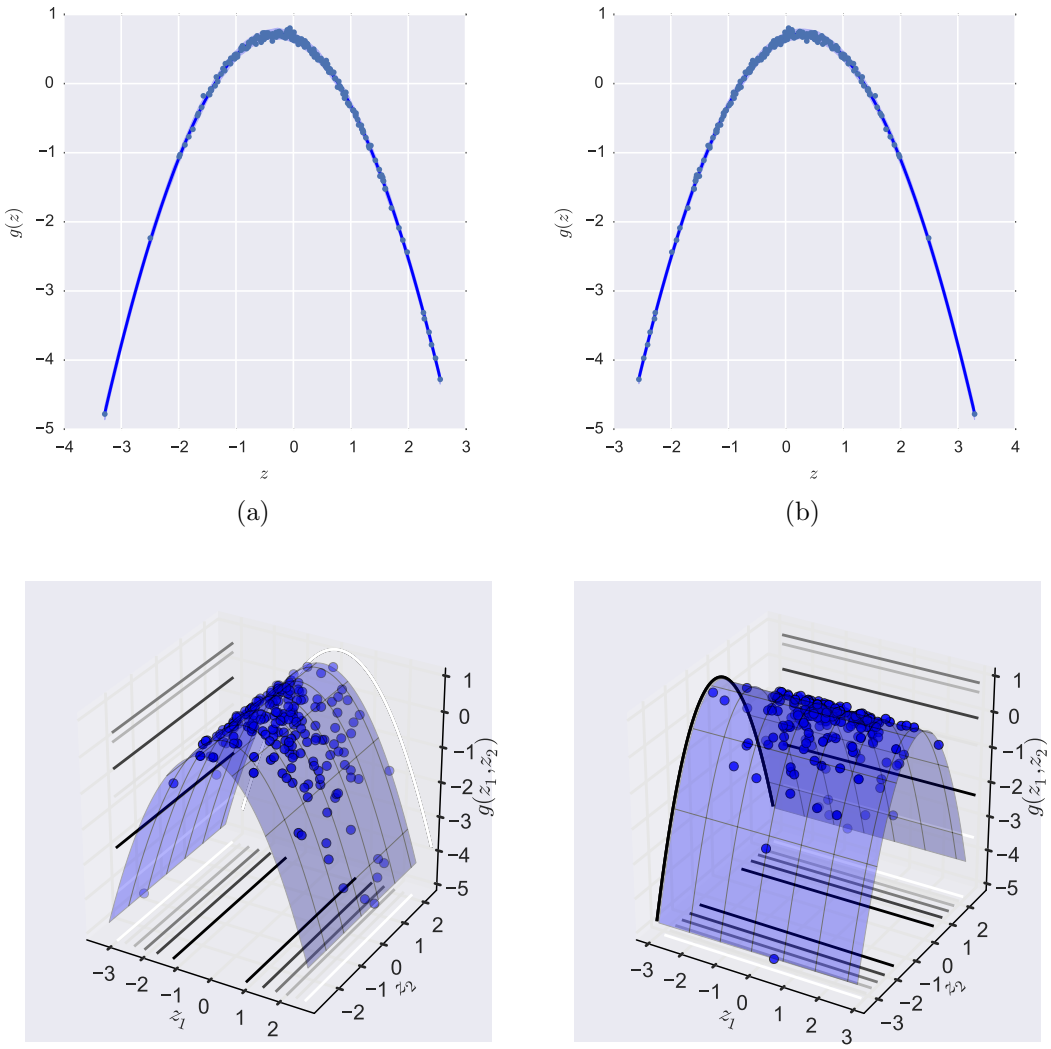


Figure 3.1. Synthetic example $d = 1$. The left and the right columns correspond to results obtained with the classic and the gradient-free approach respectively. The first and second rows depict the predictions of each method for the link function assuming a 1D and 2D underlying AS, respectively, along with a scatter plot of the projections of 60 validation inputs vs the validation outputs.

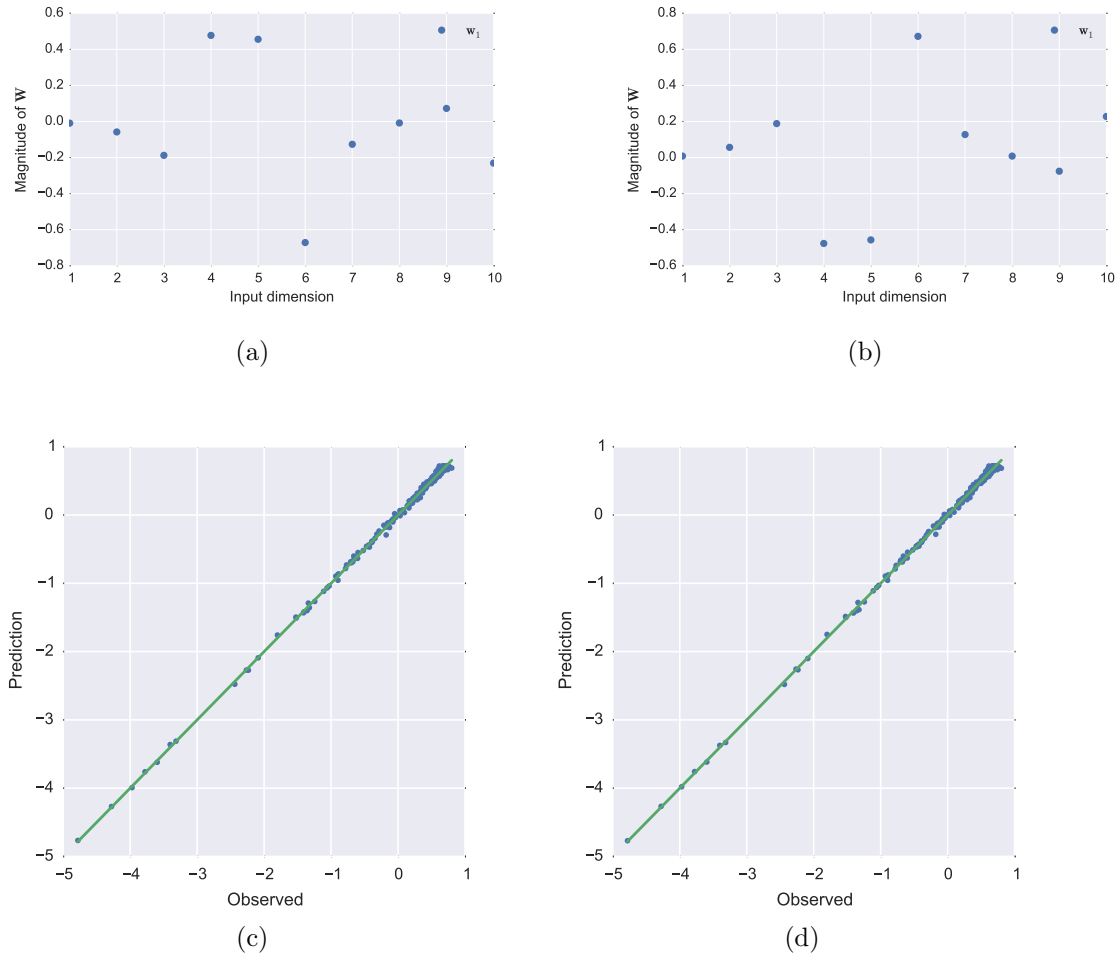


Figure 3.2. Synthetic example $d = 1$(Contd.). The left and the right columns correspond to results obtained with the classic and the gradient-free approach respectively. The first row visualizes the components of the projection matrix that each method discovers. The second row shows the observations vs model predictions for the test inputs corresponding to the surrogate for the 1d active subspace model.

3.1.1 Synthetic response with 1D active subspace

In this example the underlying AS is 1D, $d = 1$. The projection matrix is:

$$\mathbf{W} = \Big(\; -0.0091, \;\; -0.0579, \;\; -0.1877, \;\; 0.4774, \;\; 0.4559, \;\; -0.6714, \;\; -0.1264, \;\; -0.0082, \;\; 0.0724, \;\; -0.2308 \; \Big)^T, \tag{3.5}$$

and the parameters of the link function of Equation (3.3) are:

$$\alpha = -0.16113, \boldsymbol{\beta} = (\; -0.97483 \;), \text{ and } \boldsymbol{\Gamma} = (\; -1.66526 \;). \tag{3.6}$$

We make $N = 140$ observations with noise variance $s_n^2 = 0.1$. In this first example, we do not make use of the automatic method for the detection of the dimensionality of the AS. Rather, we use the plain vanilla version of both the classic and the gradient-free approaches assuming a 1D or a 2D AS. Fig. 3.1 compares the results obtained in this way. Note that the 60 validation input/outputs used in the figure were not in the training process. The quantitative agreement between the two becomes obvious once one recalls that the representation of Equation (2.7) is arbitrary up to permutations and reflections of the reduced dimensions. It is worth mentioning that for the case of a hypothetical 2D AS, both methods discovered one completely flat direction.

3.1.2 Synthetic response with 2D active subspace

In this example the underlying AS is 2D, $d = 2$. The projection matrix is:

$$\mathbf{W} = \begin{pmatrix} 0.00840 & -0.18426 & 0.34300 & -0.05347 & 0.08108 & 0.06556 & -0.41219 & 0.65424 & 0.48483 & 0.03966 \\ 0.0672 & -0.4148 & 0.4821 & 0.0755 & 0.2101 & 0.5375 & 0.0781 & -0.2002 & -0.2912 & 0.3480 \end{pmatrix}^T, \tag{3.7}$$

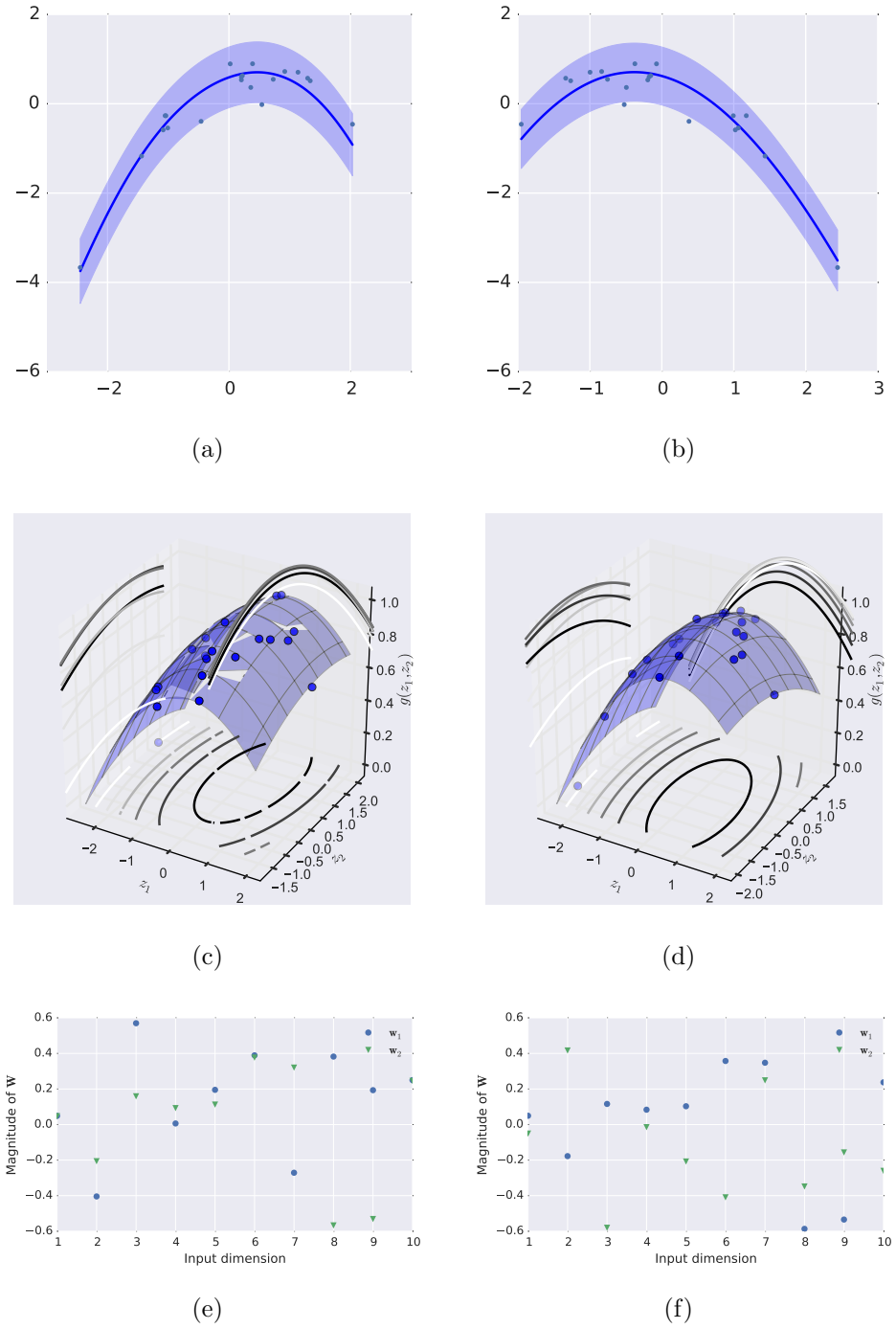
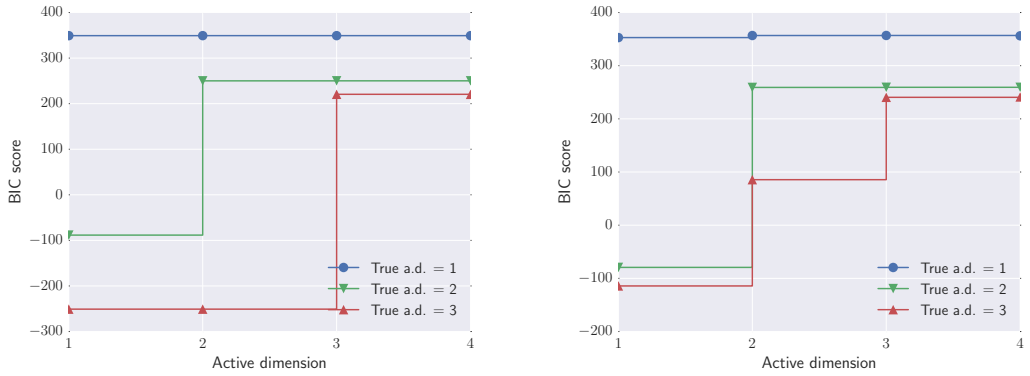


Figure 3.3. Synthetic example $d = 2$. The left and the right columns correspond to results obtained with the classic and the gradient-free approach respectively. The first row depicts the predictions of each method for the link function assuming a 2D underlying AS, along with a scatter plot of the projections of the 60 validation inputs vs the validation outputs. The second row visualizes the projection matrix that each method discovers.



(a) (b)

Figure 3.4. Synthetic example. BIC score as a function of the hypothesized active dimension for classic model (a) and the gradient-free model (b). The different lines correspond to cases with a 1D (blue, true response as in Sec. 3.1.1), 2D (green, true response as in Sec. 3.1.2, and 3D (red, true response as in details on the accompanying website) true AS.

and the parameters of the link function of Equation (3.3) are:

$$\alpha = -0.06976, \boldsymbol{\beta} = \begin{pmatrix} 0.43759 \\ 0.98696041 \end{pmatrix}, \text{ and } \boldsymbol{\Gamma} = \begin{pmatrix} -0.92567723 & -0.38399783 \\ -0.41740642 & -0.67655046 \end{pmatrix}. \tag{3.8}$$

As in Sec. 3.1.1, we make $N = 140$ observations with noise variance $s_n^2 = 0.1$, Again, we do not make use of the automatic method for the detection of the dimensionality of the AS, but assume the right 2D AS Fig. 3.3 depicts the results. Just like before, the 60 validation input/outputs used in the figure were not used in the training process. The quantitative agreement between the two approaches up to permutations and reflections of the AS is also graphically obvious.

3.1.3 Validation of BIC for the identification of the active subspace dimension

Here, we verify the effectiveness of the BIC from Sec. 2.3.3, to automatically determine the dimensionality of the AS for both the classic and the gradient-free approach.

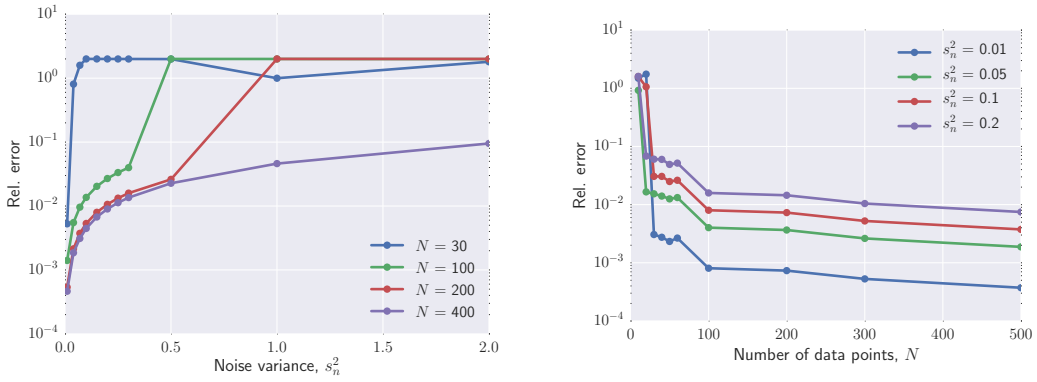
The hypothesis is that the BIC as a function of the hypothesized dimensionality of AS should become relatively flat after the hypothesized dimensionality exceeds the true AS dimensionality. This is confirmed numerically in Fig. 3.4 for the cases of a 1D, 2D, and 3D true AS. Note that for the 1D and 2D examples, the observations we used to train the models were the same as in Sec. 3.1.1 and Sec. 3.1.2. For the 3D true AS case also has an underlying response surfaces with randomly generated $\alpha, \boldsymbol{\beta}, \boldsymbol{\Gamma}$, and $\mathbf{W}$. The values used can be found in the accompanying website As before, we used $N = 140$ observations.

3.1.4 Validation of robustness to measurement noise

We conclude this subsection with a study of the robustness of the proposed scheme to measurement noise. To avoid the non-uniqueness issues mentioned earlier, we work with the 10D-input-1D-AS response surface of Sec. 3.1.1. In this case, the arbitrariness can be removed by making sure that the signs of the estimated and the true projection matrix match. We want to quantify the ability of the model to discover the true AS and how this is affected by changes in the measurement noise, s_n, as well as in the number of available observations N. A good measure of this ability is the relative error in the estimation of the projection matrix:

$$\epsilon_{\text{rel}}(s_n, N) = \frac{\|\mathbf{W}(s_n, N) - \mathbf{W}\|}{\|\mathbf{W}\|}, \tag{3.9}$$

where $\|\cdot\|$ is the Frobenius norm, $\mathbf{W}(s_n, N)$ is the estimated projection matrix when N measurements contaminated with zero mean Gaussian noise of variance s_n^2 are used, and $\mathbf{W}$ is the true projection matrix given in Equation (3.5). The results of our analysis are presented in Fig. 3.5. Fig. 3.5(a) plots the relative error, ϵ_{rel}, as a function of s_n^2 for $N = 30, 100, 200$, and 500. As expected, we observe that ϵ_{rel} increases as a function of s_n^2 and that a larger N is required to maintain a given accuracy. Fig. 3.5(b) plots the relative error, ϵ_{rel}, as a function of N for $s_n = 0.01, 0.05, 0.1$ and 0.2. We



(a) (b)

Figure 3.5. Synthetic example. Robustness of the proposed approach to measurement noise. The figure shows the evolution of the relative error in the determination of the true active subspace as a function of the measurement noise variance (keeping the number of observations constant) (a) and as a function of the number of observations (keeping the measurement noise variance constant (b)).

note that the method converges to the right answer as N increases, albeit the rate of convergence decreases for higher noise.

3.2 Elliptic Partial Differential Equation

Consider the elliptic partial differential equation [67]:

$$\nabla \cdot (c(\mathbf{s})\nabla u(\mathbf{s})) = 1,\ \mathbf{s} \in \Omega = [0,1]^2, \tag{3.10}$$

with boundary conditions

$$u(\mathbf{s}) = 0,\ \mathbf{s} \in \Gamma_1, \tag{3.11}$$

$$\nabla u(\mathbf{s}) \cdot \mathbf{n} = 0,\ \mathbf{s} \in \Gamma_2, \tag{3.12}$$

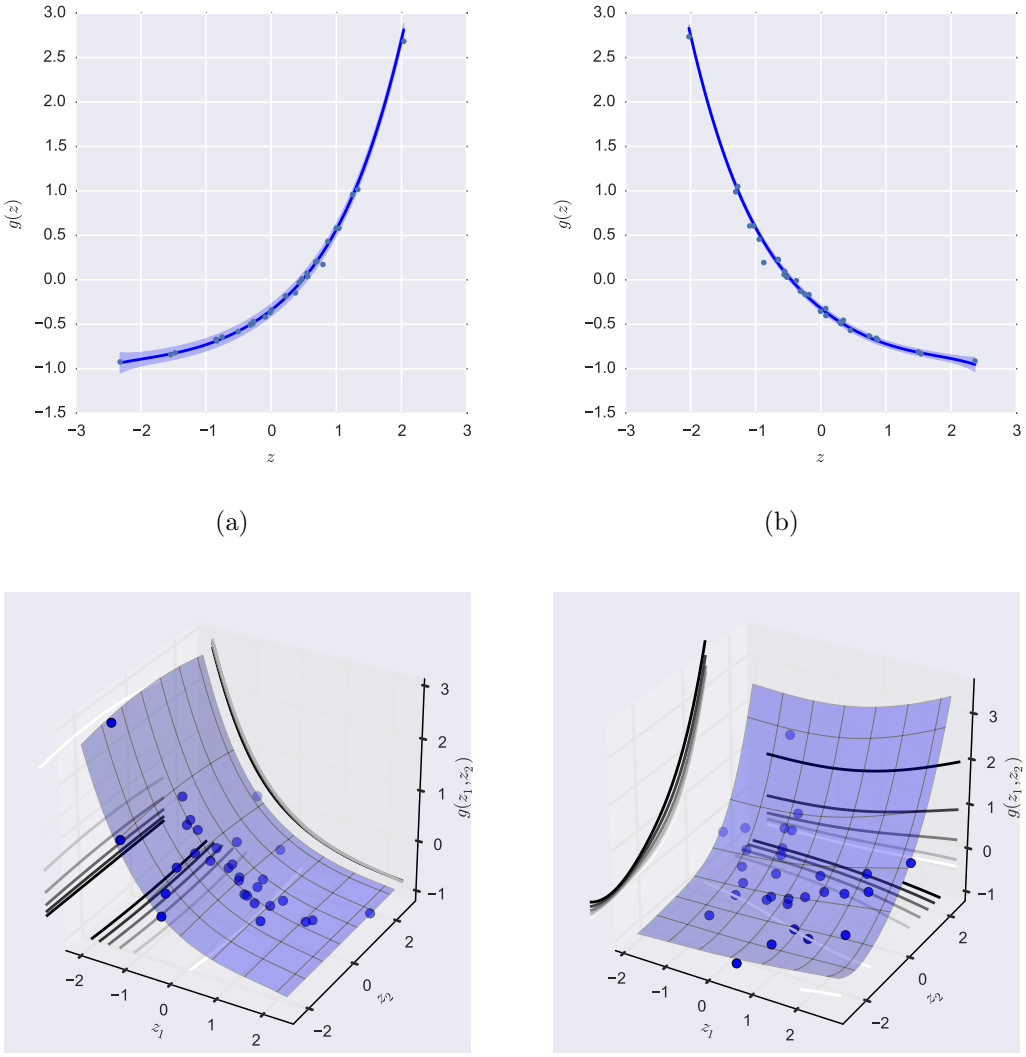
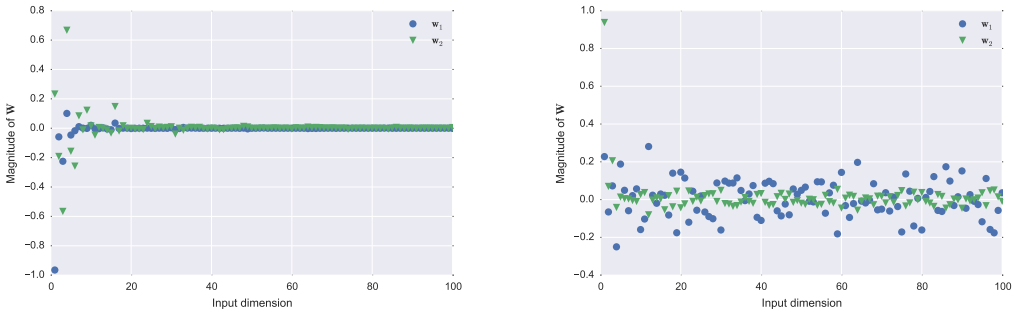


Figure 3.6. Elliptic PDE, long correlation length ($\ell = 1$). The left and the right columns correspond to results obtained with the classic and the gradient-free approach respectively. The first and second rows depict the predictions of each method for the link function assuming a 1D and 2D underlying AS, respectively, along with a scatter plot of the projections of 30 validation inputs vs the validation outputs. The third row visualizes the projection matrix that each method discovers.



(f) (g)

Figure 3.6. (Continued).

where Γ_1 is the top, bottom and left boundaries and Γ_2 denotes the right boundary of Ω. We assume that the conductivity field is unknown and model its logarithm as a Gaussian random field with an exponential correlation function:

$$C(\mathbf{s}, \mathbf{s}'; \ell) = \exp\left\{-\frac{|s_1 - s_1'| + |s_2 - s_2'|}{\ell}\right\}, \tag{3.13}$$

with correlation length $\ell > 0$. Using a truncated Karhunen-Loève expansion (KLE), the logarithm of the conductivity can be expressed as:

$$\log \mathbf{c}(\mathbf{s}; \mathbf{x}) := \sum_{i=1}^{100} x_i \sqrt{\lambda_i} \phi_i(\mathbf{s}), \tag{3.14}$$

where λ_i and $\phi_i(\mathbf{s})$ are the eigenvalues and eigenfunctions of the correlation function, Equation (3.13), and $\mathbf{x}$ is a random vector modeled as uniformly distributed on $[-1, 1]^{100}$, i.e., $\mathbf{x} \sim \mathcal{U}([-1, 1]^{100})$. The latter violates the theoretical form of the KLE, but guarantees the existence of a solution to the boundary value problem defined by Equation (3.10)-Equation (3.12) for all $\mathbf{x}$. Given any value for $\mathbf{x}$, the solution of the boundary value problem is $u(\cdot; \mathbf{x})$.

In our analysis, we attempt to learn the following scalar quantity of interest:

$$f(\mathbf{x}) := f[u(\cdot; \mathbf{x})] := \frac{1}{|\Gamma_2|} \int_{\Gamma_2} u(\mathbf{s}; \mathbf{x}) ds, \tag{3.15}$$

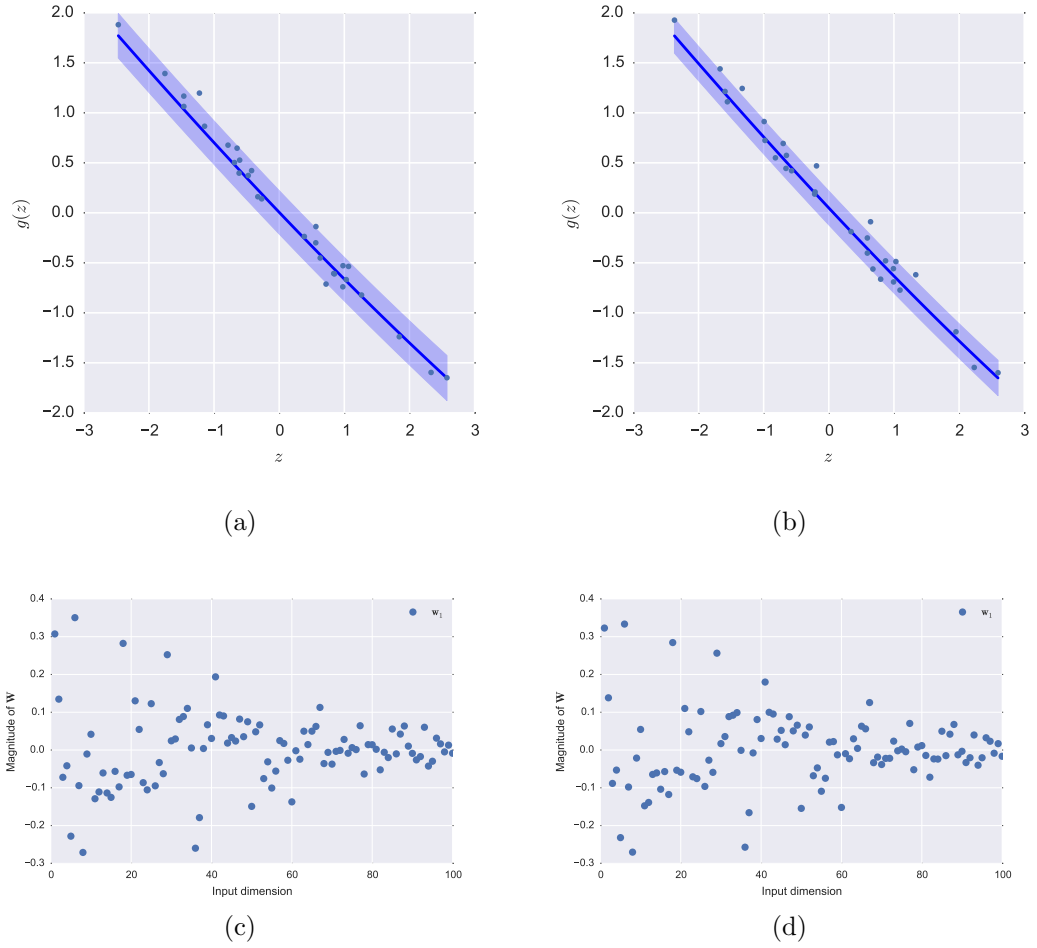
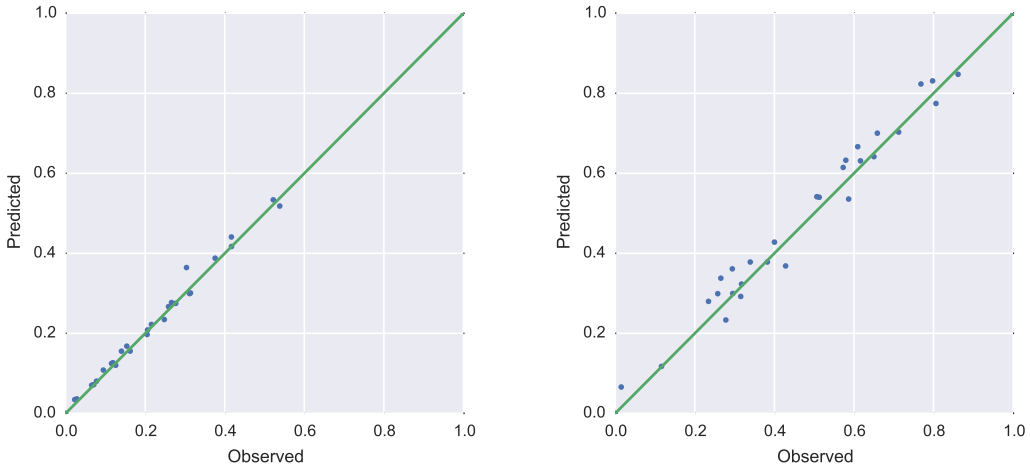
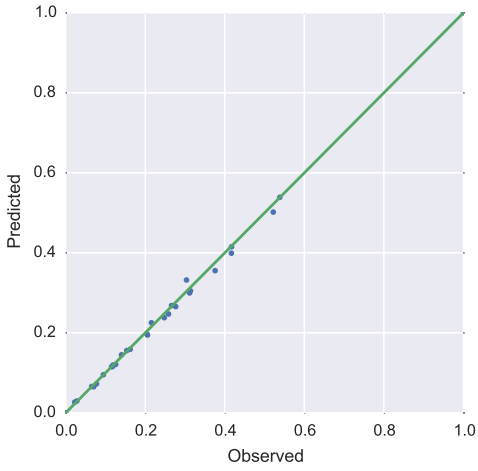


Figure 3.7. Elliptic PDE, long correlation length ($\ell = 0.01$). The left and the right columns correspond to results obtained with the classic and the gradient-free approach respectively. The first and second rows depict the predictions of each method for the link function assuming a 1D and 2D underlying AS, respectively, along with a scatter plot of the projections of 30 validation inputs vs the validation outputs. The third row visualizes the projection matrix that each method discovers.

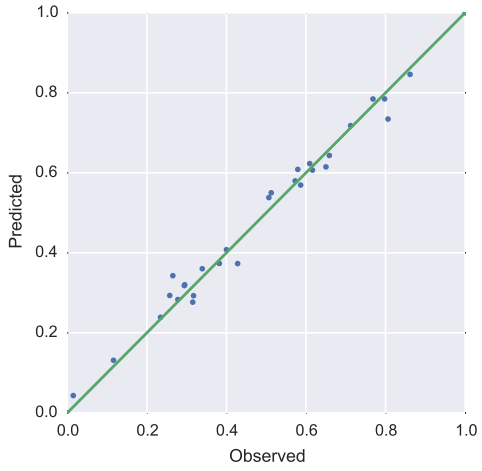


(a)

(b)



(c)



(d)

Figure 3.8. Elliptic PDE. The dots correspond to true observed responses vs predicted ones for 30 validation inputs for the long ($\ell = 1$, left) and short ($\ell = 0.01$, right) correlation cases. Perfect predictions would fall on the green 45° line of each subplot. The top row corresponds to the gradient-free approach while the bottom row corresponds to the classic approach.

using both the classic, Sec. 2.2, and the gradient-free approach, Sec. 2.3. We examine two cases exhibiting two different correlation lengths. The first case uses a long correlation length, $\ell = 1$, and the second case a short correlation length $\ell = 0.01$. In both cases, we use $N = 270$ noiseless observations of input-output pairs for training purposes, while setting 30 aside for validation. The data along with the `MATLAB` code that generates them, developed by Paul Constantine, can be obtained from https://bitbucket.org/paulcon/active-subspace-methods-in-theory-and-practice/src.

Fig. 3.6 shows the results we obtain using the long, $\ell = 1$ correlation length. The first and second rows of this figure depict the discovered link function under the assumption of a 1D and a 2D AS, respectively. Note that both methodologies agree on the most important AS dimension, but slightly disagree on the second, albeit relatively flat, dimension. A close examination of the discovered projection matrices, third line of the figure, reveals the following. The most important column of the classic projection matrix, $\mathbf{w}_1$, is matched with the negative of the second column discovered by the gradient free approach, $-\mathbf{w}_2$. The latter, however, looks like a "noisy" version of the former. This is reasonable if one takes into account that the gradient-free approach uses significantly less information than the classic approach. Finally, we notice that the columns of secondary importance do not match. This discrepancy is unimportant given that the BIC score eventually selects a 1D AS.

Fig. 3.7 shows the results for the more challenging case of the short correlation length. We present the 1D representation of the link function, as discovered by the classic and the gradient-free approach, in the first row and show the components of the projection matrix estimated by each methodology in the second row. We note that both methodologies show similar 1D active subspace representation of the surrogate. Indeed, this is the most important dimension as the response should be flat along the 2^{nd} dimension. We find that the components of projection matrix estimated by both methods are in qualitative agreement for the 1D surrogate. The BIC score approximately increases by 12% between the 1D and 2D active subspace model provided by the gradient-free approach. Thus, the quality of the 2D response

surface learned by the gradient-free approach suffers from lack of sufficient data. Given more observations, it is expected that the response surface will qualitatively converge to one which is flat along the 2^{nd} active dimension.

Table 3.1. BIC Score for $\ell = 1, 0.01$ corresponding to classic and gradient-free methodologies.

	$\ell = 1$	$\ell = 0.01$
Classic approach	1.87×10^{-5}	3.2×10^{-6}
Gradient-free approach	2.66×10^{-5}	3.57×10^{-6}

Fig. 3.8 shows the comparison between the prediction on the test inputs and the actual response. The closer the points lie to the green 45° line, the more accurate the prediction. We make these comparisons for the 1D representation of the link function for both the short and long correlation length cases. It appears that the predictive capabilities of the classic approach are slightly better than the gradient-free approach. A comparison of the RMS error for the predictions by each methodology confirms this although the difference is essentially negligible given its order of magnitude. We tabulate this data in Table 3.2.

Table 3.2. Predictive RMS errors for $\ell = 1, 0.01$ corresponding to classic and gradient-free methodologies.

	$\ell = 1$	$\ell = 0.01$
Classic approach	1.87×10^{-5}	3.2×10^{-6}
Gradient-free approach	2.66×10^{-5}	3.57×10^{-6}

3.3 Granular Crystals

One dimensional granular chains have attracted significant attention[references] owing to their unique and highly non-linear dynamical properties. These chains

support the formation and propagation of highly localized, elastic stress waves when perturbed with a small to moderate external excitation [97–101]. Granular crystals are typically described by a fully elastic model know as the Hertz contact model [102]. We now proceed to setup the problem.

Consider a one-dimensional chain of n_p particles whose displacements from the equilibrium positions are described by the position vector $\mathbf{q} = \left(q_1, q_2, \cdots, q_{n_p}\right)$. Each bead has a radius R_i and Young's modulus E_i. The number of contact points is given by $s = n_p - 1$. The density of the particles is a constant $\rho = 7,900\text{kg/m}^3$ and the mass of each particle is $m_i = \rho\frac{4}{3}\pi R_i^3$. In the current problem we leave no gap between the particles. We excite this system by striking the n_p-th particle with a striker traveling at velocity v_s.

Define the parameter vector for the system, $\mathbf{x}$, as follows:

$$\mathbf{x} = \left(R_1, R_2, \ldots, R_{n_p}, E_1, E_2, \ldots, E_{n_p}, v_s\right). \tag{3.16}$$

The displacement vector satisfies Newton's law of motion:

$$m_i(\mathbf{x})\ddot{q}_i = F_i\left(\mathbf{q}; \mathbf{x}\right), \tag{3.17}$$

with initial conditions:

$$\begin{aligned} q_i(0) &= 0, \\ \dot{q}_i(0) &= 0, \ \forall\, i = \{1, 2, 3, \cdots, n_p - 1\}, \\ \dot{q}_{n_p}(0) &= -v_s. \end{aligned}$$

Let $q(t; \mathbf{x})$ be the solution to this initial value problem. We are interested in characterizing the properties of force waves propagated through the granular crystal. To

this end, we will be observing the force on each particle as a function of time for a given set of parameters $\mathbf{x}$.

$$F_i\left(t;\mathbf{x}\right) \equiv F_i\left(\mathbf{q}\left(t;\mathbf{x}\right);\mathbf{x}\right). \tag{3.18}$$

That is, for each $\mathbf{x}$, we obtain, by integrating the equations of motion, the force at a finite number of timesteps, $0 = t_1 < \cdots < t_{n_t}$, say $\mathbf{F}(\mathbf{x}) := \{F_i(t_j;\mathbf{x}) : i = 1,\ldots,n_p, j = 1,\ldots,n_t\}$. The dimensionality of the output is given by the number of time steps times the number of particles. This is a very high-dimensional output and we first reduce it's dimensionality by fitting the output to a solitary wave whose properties-amplitude, wave width and velocity, are particle dependent. Let the properties of the soliton over the i-th particle be amplitude $A_i\left(\mathbf{x}\right)$, wave speed $V_i\left(\mathbf{x}\right)$ and width $W_i\left(\mathbf{x}\right)$.

Then, the force on the i-th particle at time step t_j is the approximated as:

$$\hat{F}_i\left(t_j;\mathbf{x}\right) = \begin{cases} A_i(\mathbf{x})\sin^4\left(\dfrac{\pi(r_i(\mathbf{x}) + V_i(\mathbf{x})t_j)}{W_i(\mathbf{x})}\right) & ,0 < \dfrac{\pi(r_i(\mathbf{x}) + V_i(\mathbf{x})t_j)}{W_i(\mathbf{x})} < \pi. \\ 0 & ,\text{otherwise.} \end{cases} \tag{3.19}$$

where $r_i(\mathbf{x})$ is the equilibrium position of particle i. Let us denote the approximated forces on all the particles at all time-steps as $\hat{\mathbf{F}}(\mathbf{x}) := \{\hat{F}_i(t_j;\mathbf{x}) : i = 1,\ldots,n_p, j = 1,\ldots,n_t\}$. The dimension of the output space is now reduced to number of solitary waves included times 3. Specifically, for a perfect chain, i.e., a chain with all parameters equal to the mean value, the dimension of the reduced output space will be 3. We now consolidate the parameters of the approximation into the following vector:

$$\mathbf{\Theta}\left(\mathbf{x}\right) = \left\{\left(A_i\left(\mathbf{x}\right), V_i\left(\mathbf{x}\right), W_i\left(\mathbf{x}\right)\right)\right\}_{i=1}^{n_p} \tag{3.20}$$

Estimation of the parameters $\mathbf{\Theta}(\mathbf{x})$ is an optimization problem for which we first define a squared-error loss function:

$$\mathcal{L}(\mathbf{\Theta}(\mathbf{x})) = \left[\hat{\mathbf{F}}(\mathbf{x}) - \mathbf{F}(\mathbf{x})\right]^2 \tag{3.21}$$

We now obtain an estimate of the vector $\mathbf{\Theta}(\mathbf{x})$ through a least squares minimization:

$$\mathbf{\Theta}^*(\mathbf{x}) = \{A_i^*(\mathbf{x}), V_i^*(\mathbf{x}), W_i^*(\mathbf{x})\}_{i=1}^{n_p} = \underset{\mathbf{\Theta}(x)}{arg\,min} \mathcal{L}(\mathbf{\Theta}(\mathbf{x})) \tag{3.22}$$

For the purpose of this experiment, we study the properties of the soliton localized over two particles - the 10^{th} and 20^{th}. Thus, we extract the following scalar quantities from $\mathbf{\Theta}$: $y_1 = A_{10}, y_2 = V_{10}, y_3 = W_{10}, y_4 = A_{20}, y_5 = V_{20}, y_6 = W_{20}$. We repeat this entire process for 1000 samples of $\mathbf{x}$ and construct the output vectors $\mathbf{y}_1, \mathbf{y}_2, \cdots, \mathbf{y}_6$ such that $\mathbf{y}_i \in \mathbb{R}^{1000}$.

We treat the vector of all the Young's moduli $\mathbf{E} = \left(E_1, E_2, \cdots, E_{n_p}\right)$ and the vector of all the particle radii $\mathbf{R} = \left(R_1, R_2, \cdots, R_{n_p}\right)$ as seperate stochastic inputs and take the velocity of the striker v_s to be constant throughout. Thus, we define the following two input matrices: $\mathbf{X}_1 = [\{\mathbf{R}_i\}_{i=1}^{1000}]$ and $\mathbf{X}_2 = [\{\mathbf{E}_i\}_{i=1}^{1000}]$, i.e. $\mathbf{X}_i \in \mathbb{R}^{1000 \times 3n_p}$.

We now proceed to apply our proposed gradient-free AS approach to build a cheap-to-evaluate surrogate for propagating uncertainty through this system. We can consider any possible combinations of data-sets $\mathcal{D}_{ij} = \{\mathbf{X}_i, \mathbf{y}_j\},\ \forall i \in \{1, 2\}\ \&\ \forall j \in \{1, 2, \cdots, 6\}$, and build the corresponding surrogates. Specifically, we show the following cases:

- **Case 1**: Input = $\mathbf{X}_2$; Output QoI = $\mathbf{y}_1$
- **Case 2**: Input = $\mathbf{X}_2$; Output QoI = $\mathbf{y}_3$
- **Case 3**: Input = $\mathbf{X}_1$; Output QoI = $\mathbf{y}_5$

We train the model on 200 observations with inputs picked uniformly within the range $(190GPa, 200GPa)$ for Young's moduli input and $(9.4mm, 9.6mm)$ for radii

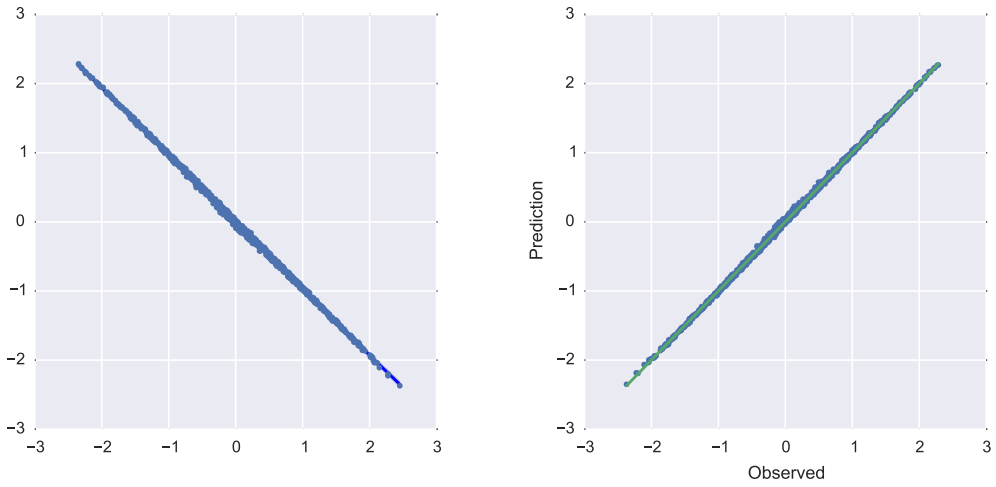
input. Note that we construct a different AS for each one of the cases. We use the remaining 800 samples to test the accuracy of the surrogate.

3.3.1 Results

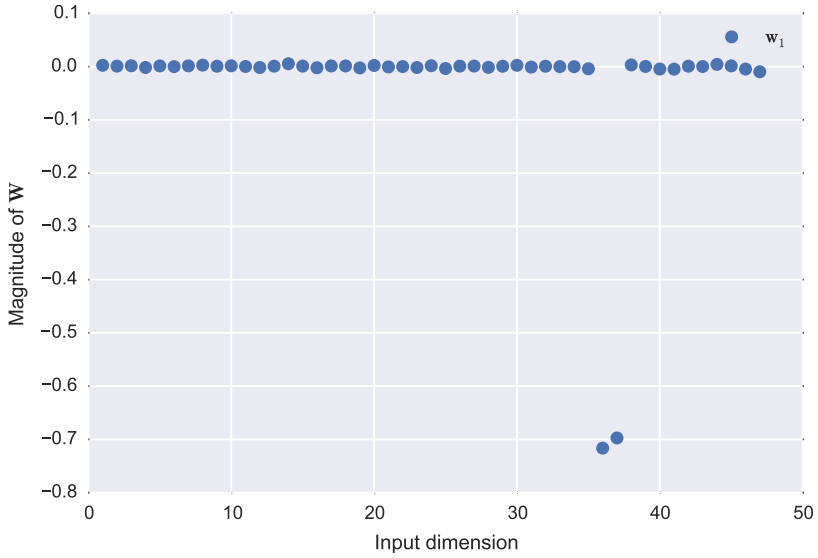
We find that most of the stochasticity of this high dimensional problem is exhibited on a one dimensional active subspace. We present plots of the AS representation of the link function, projection matrix and predictions vs observations plots for a few cases. We note that the underlying response surfaces obtained are approximately linear. There is very good agreement between the training data-set output predictions as compared to the actual training set outputs. On taking a closer look at the projection matrices we find that the weights assigned to the particles after the particle in consideration are approximately zero. From Fig. 3.9, we find that the output(amplitude of the soliton corresponding to particle 10) is almost entirely dependent on the Young's modulus corresponding to the 10^{th} particle. From Fig. 3.10, we find that the projection matrix has positive weights corresponding to the first 9 particles and negative weights upto a few particles after that. And finally, from Fig. 3.11 we note that the effect of the particle radii on the velocity of the soliton corresponding to particle 20 is largely localized around the 20^{th} particle.

3.3.2 Uncertainty Quantification

Having built a cheap-to-evaluate response GP response surface, we now demonstrate a few examples of Uncertainty Propagation. We assign a normal distribution to the inputs with mean $\mu = \frac{a+b}{2}$ and variance $\sigma^2 = 1\%$ of μ, where a, b are the minimum and maximum values of the observed input. Then, we sample 100000 observations of the input from this distribution and use the surrogate to generate predictions. Finally, we plot the marginal and joint distributions of some of the outputs as shown

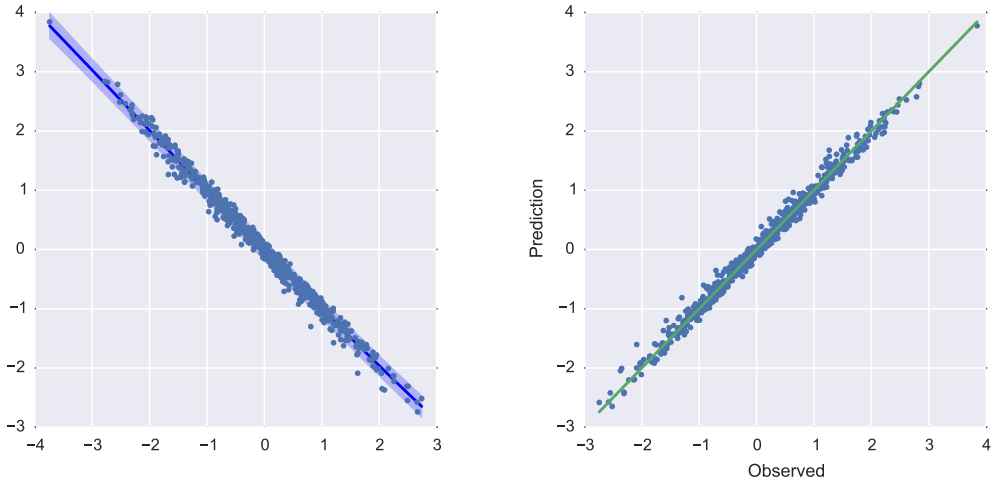


(a) (b)

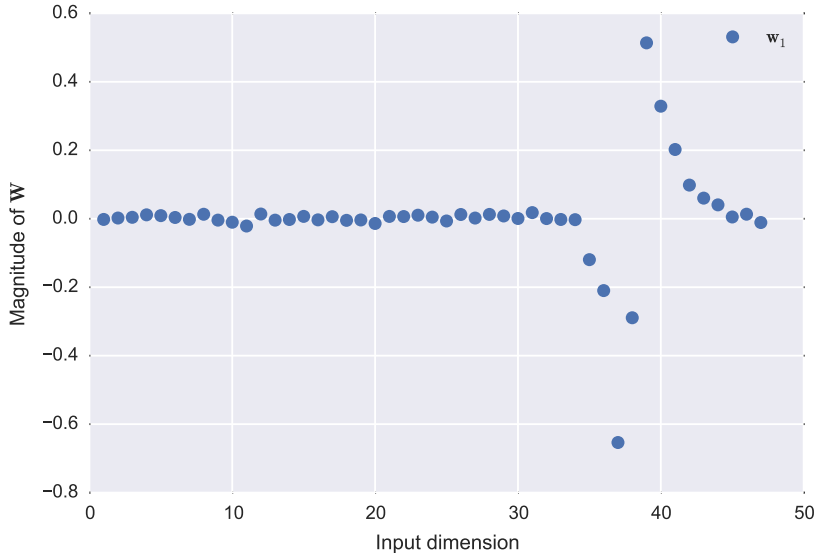


(c)

Figure 3.9. Plots for the 10^{th} particle corresponding to Young's modulus input and soliton amplitude output. The first plot shows the response surface in the active subspace. the second plot depicts the test observations vs model prediction plot. The final plot depicts the components of the projection matrix.



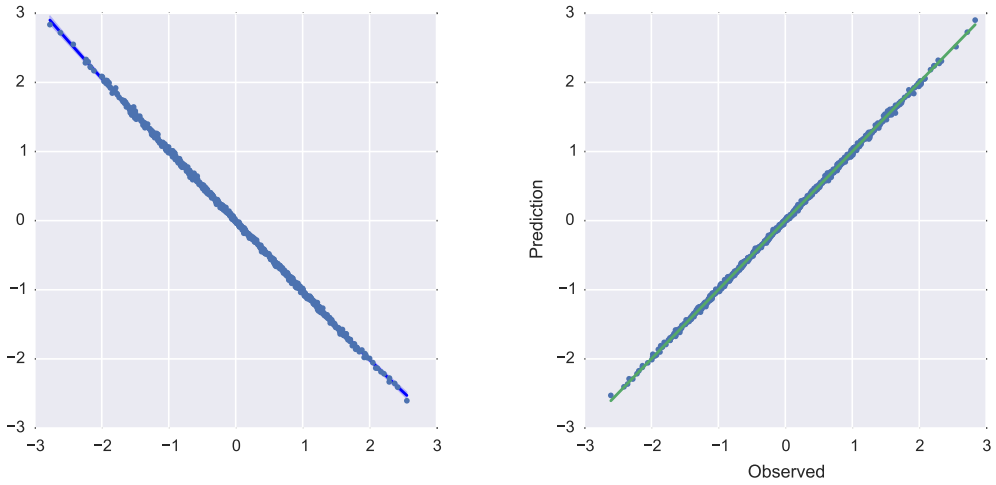
(a) (b)



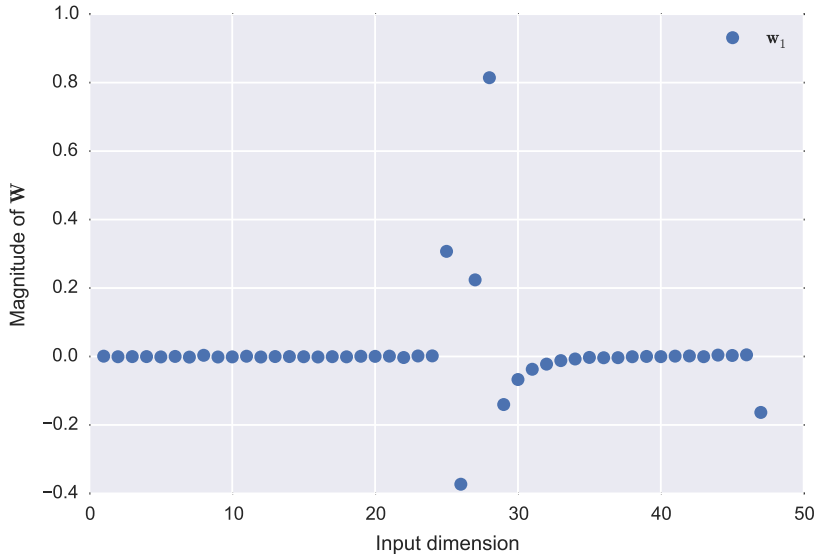
(c)

Figure 3.10. Plots for the 10^{th} particle corresponding to Young's modulus input and soliton wave width output. The first plot shows the response surface in the active subspace. the second plot depicts the test observations vs model prediction plot. The final plot depicts the components of the projection matrix.

in Fig. 3.12, 3.13 and 3.14. Note that for a normal input distribution, the output is also approximately normal.

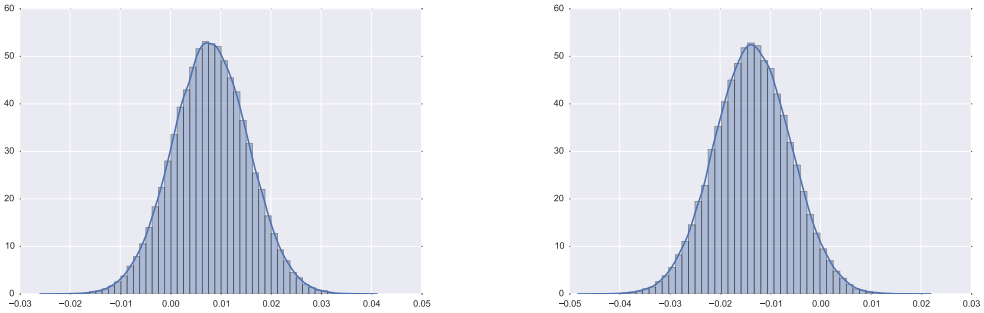


(a) (b)



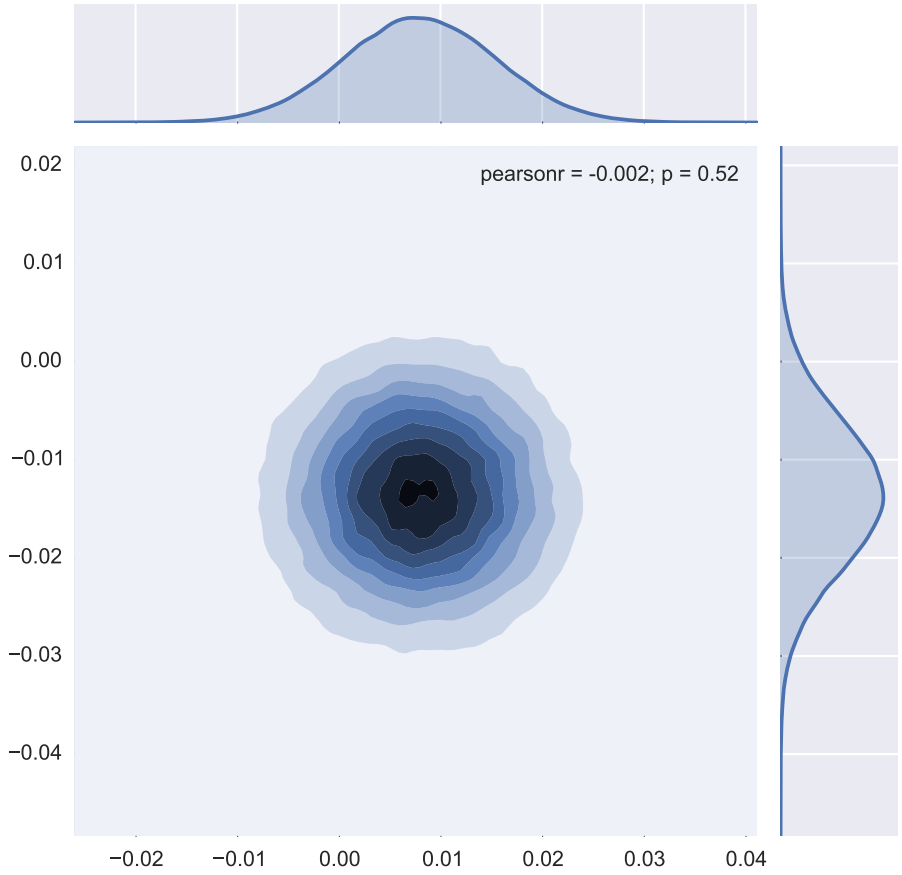
(c)

Figure 3.11. Plots for the 20^{th} particle corresponding to particle radii input and soliton wave velocity output. The first plot shows the response surface in the active subspace. the second plot depicts the test observations vs model prediction plot. The final plot depicts the components of the projection matrix.



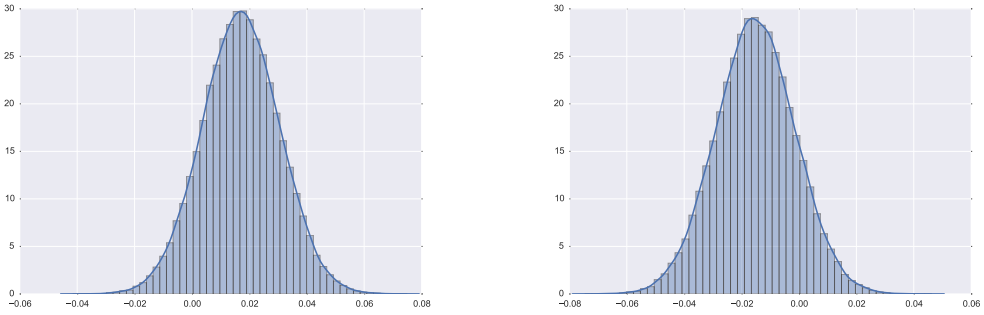
(a)

(b)

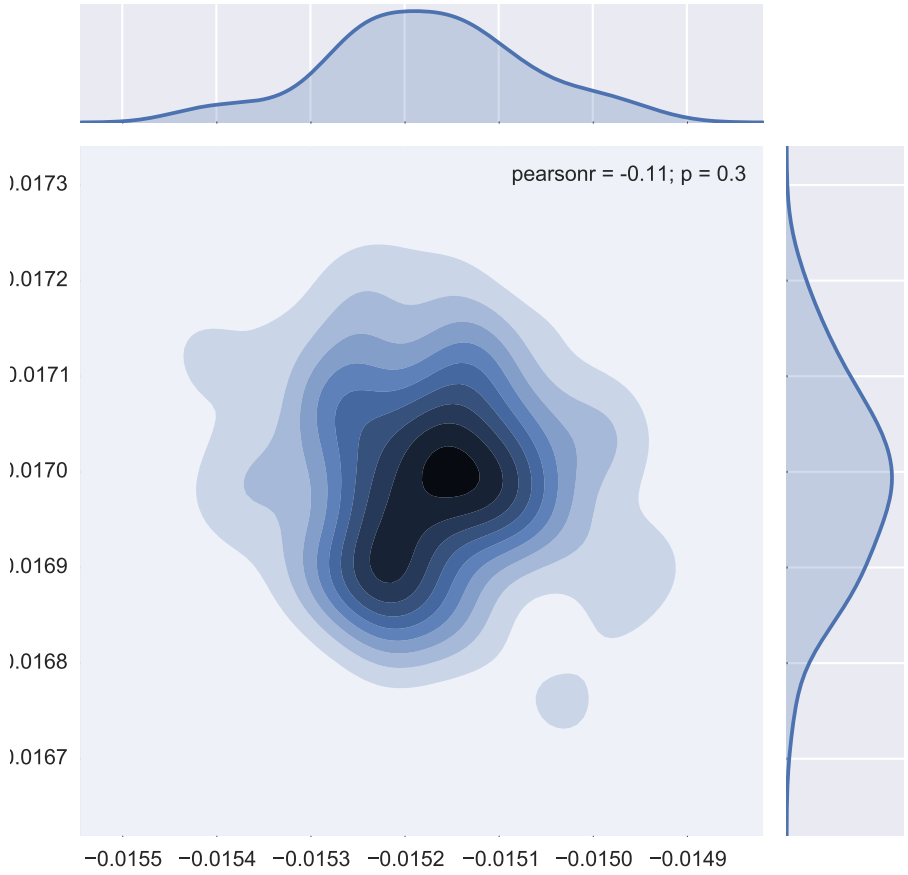


(c)

Figure 3.12. Propagating the uncertainty by assigning a normal distribution to the Young's moduli. (a) Marginal distribution of the velocity of the soliton over the 10^{th} particle; (b) Marginal distribution of the width of the soliton over the 10^{th} particle; (c) Joint distribution of the velocity and width of the soliton over the 10^{th} particle.

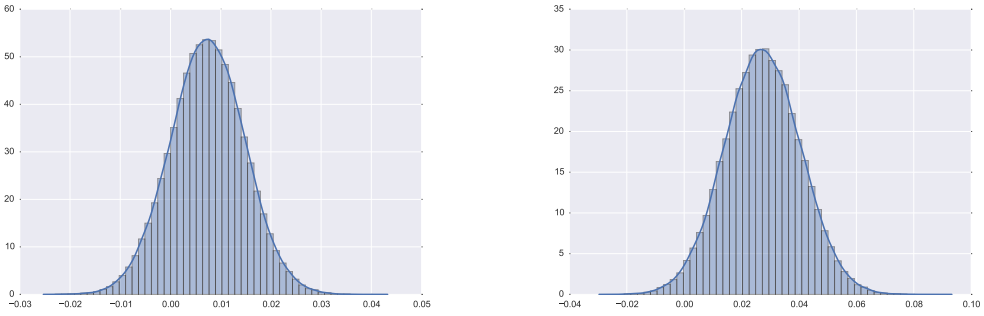


(a) (b)

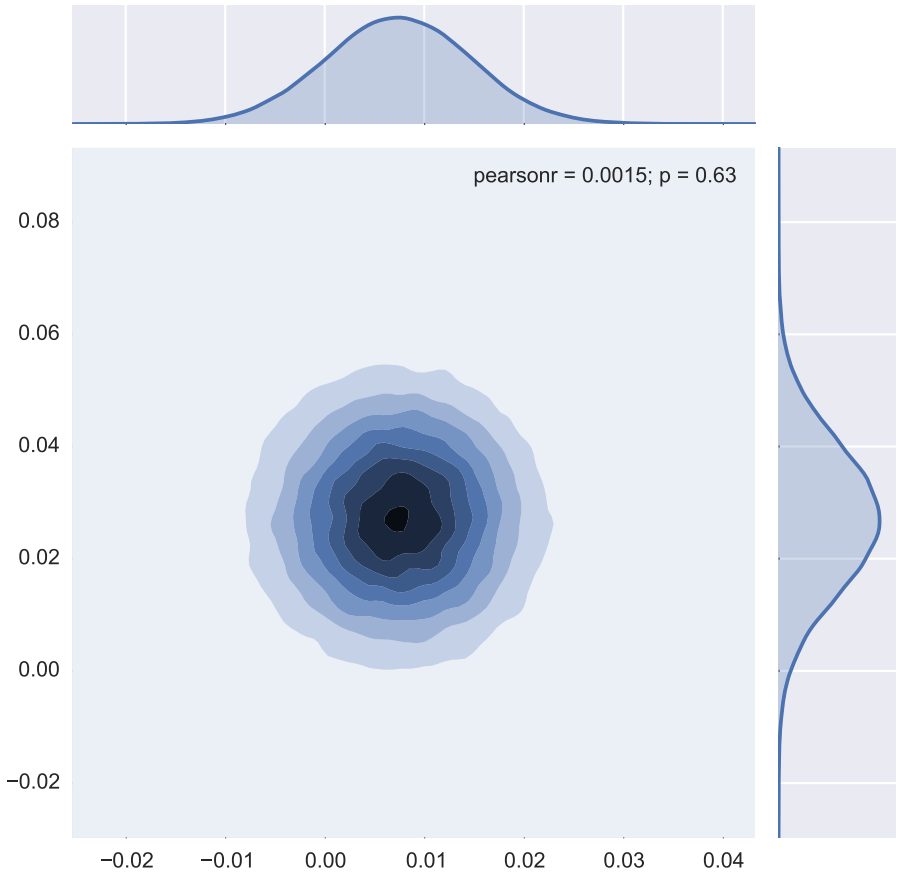


(c)

Figure 3.13. Propagating the uncertainty by assigning a normal distribution to the radii. (a) Marginal distribution of the amplitude of the soliton over the 20^{th} particle; (b) Marginal distribution of the velocity of the soliton over the 20^{th} particle; (c) Joint distribution of the velocity and amplitude of the soliton over the 20^{th} particle.



(a) (b)



(c)

Figure 3.14. Propagating the uncertainty by assigning a normal distribution to both the radii and Young's moduli. (a) Marginal distribution of the amplitude of the soliton over the 10^{th} particle. for the given distribution of the Young's moduli; (b) Marginal distribution of the width of the soliton over the 20^{th} particle for the given distribution of the radii; (c) Joint distribution of the width of the soliton over the 20^{th} particle and the amplitude of the soliton over the 10^{th} particle.

4. SUMMARY

We have developed a gradient-free approach to active subspace (AS) discovery and exploitation suitable for dealing with noisy outputs. We did so by developing a novel Gaussian process regression model with built-in dimensionality reduction. Specifically we represented the AS as an orthogonal projection matrix that constitutes a hyper-parameter of the covariance function to be estimated from the data by maximizing the likelihood. Towards this end, we devised a two-step optimization procedure that ensures the orthogonality of the projection matrix by exploiting recent results on the description of the Stiefel manifolds. An addendum of the probabilistic approach is the ability to use the Bayesian information criterion (BIC) to automatically select the dimensionality of the AS. We validated our method using both synthetic examples with known AS and by comparing directly to the classic gradient-based AS approach. Finally, we used our method to study the effect of geometric and material uncertainties in force waves propagated through granular crystals.

This work is a first step towards a fully Bayesian AS-based surrogate, a persistent theme of our current research plans. As argued in [38], Bayesian surrogates should be capable of quantifying all the epistemic uncertainty induced by limited data, since quantification of this epistemic uncertainty is the key to deriving problem-specific information acquisition policies, i.e., rules for deciding where to sample the model next in order to obtain the maximum amount of information towards a specific task. A fully Bayesian treatment requires the specification of priors for all the hyper-parameters of the covariance function and the derivation of Markov Chain Monte Carlo (MCMC) schemes to sample from the posterior of the model. The big challenge is the construction of proposals that force $\mathbf{W}$ to remain on the Stiefel manifold, which could be achieved, for example, by modifying the Riemann manifold Hamiltonian MC of [103].

Such approaches would open the way for more robust AS dimensionality selection, e.g., by reversible-jump MC [104] or by directly computing the model evidence.

Many physical models do not have an AS. They may have, however, a non-linear low-dimensional manifold exhibiting maximal response variability. Assuming that this low-dimensional manifold is a Riemann manifold, i.e., locally isomorphic to a Eucledian space, a potential approach would be to consider mixtures of the model proposed in this work. To this end, the results of [37] on infinite mixtures of GP's could be leveraged. The latter is also a subject of on-going research.

LIST OF REFERENCES

LIST OF REFERENCES

[1] R.C. Smith. *Uncertainty quantification : theory, implementation, and applications*. Philadelphia : SIAM, 2014.

[2] W. Chen. Efficient uncertainty analysis methods for multidisciplinary robust design. *AIAA Journal*, 40(3):545–552, 2002.

[3] F. D'Auria and G.M. Galassi. Outline of the uncertainty methodology based on accuracy extrapolation. *Nuclear Technology*, 109(1), 1995.

[4] W.L. Oberkampf, J.C. Helton, C.A. Joslyn, S.F. Wojtkiewicz, and S. Ferson. Challenge problems: uncertainty in system response given uncertain parameters. *Reliability Engineering & System Safety*, 85(1):11–19, 2004.

[5] J.S. Liu. *Monte Carlo strategies in scientific computing*. New York : Springer, New York, 2001.

[6] C.P. Robert. *Monte Carlo statistical methods*. New York : Springer, New York, 2nd ed. edition, 2004.

[7] W.J. Morokoff and R.E. Caflisch. Quasi-Monte Carlo integration. *Journal of Computational Physics*, 122(2):218–230, 1995.

[8] A. Barth, C. Schwab, and N. Zollinger. Multi-level Monte Carlo finite element method for elliptic PDEs with stochastic coefficients. *Numerische Mathematik*, 119(1):123–161, 2011.

[9] F.Y. Kuo, C. Schwab, and I.H. Sloan. Quasi-Monte Carlo finite element methods for a class of elliptic partial differential equations with random coefficients. *SIAM Journal on Numerical Analysis*, 50(6):3351–3374, 2012.

[10] A. Tarantola. *Inverse problem theory and methods for model parameter estimation*. SIAM, 2004.

[11] I. Bilionis, B.A. Drewniak, and E.M. Constantinescu. Crop physiology calibration in the clm. *Geoscientific Model Development*, 8(4):1071–1083, 2015. GMD http://www.geosci-model-dev.net/8/1071/2015/gmd-8-1071-2015.pdf.

[12] J.C. Spall. *Introduction to stochastic optimization*. Wiley, 2003.

[13] I. Bilionis and P.S. Koutsourelakis. Free energy computations by minimization of kullback-leibler divergence: An efficient adaptive biasing potential method for sparse representations. *Journal of Computational Physics*, 231(9):3849–3870, 2012.

[14] I. Bilionis and N. Zabaras. A stochastic optimization approach to coarse-graining using a relative-entropy framework. *Journal of Chemical Physics*, 138(4), 2013.

[15] A. O'Hagan. Monte carlo is fundamentally unsound. *Statistician*, 36(2-3):247–249, 1987. J4389 Times Cited:13 Cited References Count:3.

[16] D. Xiu and G.E. Karniadakis. The wiener–askey polynomial chaos for stochastic differential equations. *SIAM journal on scientific computing*, 24(2):619–644, 2002.

[17] D. Xiu and J. Hesthaven. High-order collocation methods for differential equations with random inputs. *SIAM Journal on Scientific Computing*, 27(3):1118–1139, 2005.

[18] D. Xiu. Efficient collocational approach for parametric uncertainty analysis. *Communications in computational physics*, 2(2):293–309, 2007.

[19] I. Babuska, F. Nobile, and R. Tempone. A stochastic collocation method for elliptic partial differential equations with random input data. *SIAM Journal on Numerical Analysis*, 45(3):1005–1034, 2007.

[20] S. Sankaran and A.L. Marsden. A stochastic collocation method for uncertainty quantification and propagation in cardiovascular simulations. *Journal of biomechanical engineering*, 133(3):031001, 2011.

[21] H.M. Gutmann. A radial basis function method for global optimization. *Journal of Global Optimization*, 19(3):201–227, 2001.

[22] R.G. Regis and C.A. Shoemaker. Constrained global optimization of expensive black box functions using radial basis functions. *Journal of Global Optimization*, 31(1):153–171, 2005.

[23] I. Bilionis and N. Zabaras. Multidimensional adaptive relevance vector machines for uncertainty quantification. *SIAM Journal on Scientific Computing*, 34(6):B881–B908, 2012.

[24] X. Ma and N. Zabaras. An adaptive hierarchical sparse grid collocation algorithm for the solution of stochastic differential equations. *Journal of Computational Physics*, 228(8):3084–3113, 2009.

[25] M.T.C. Currin, M. Morris, and D. Ylvisaker. A bayesian approach to the design and analysis of computer experiments. Report, Oak Ridge Laboratory, 1988.

[26] J. Sacks, W.J. Welch, T. Mitchell, and H.P. Wynn. Design and analysis of computer experiments. *Statistical Science*, 4(4):409–423, 1989.

[27] C. Currin, T. Mitchell, M. Morris, and D. Ylvisaker. Bayesian prediction of deterministic functions, with applications to the design and analysis of computer experiments. *Journal of the American Statistical Association*, 86(416):953–963, 1991.

[28] I. Bilionis and N. Zabaras. Multi-output local gaussian process regression: Applications to uncertainty quantification. *Journal of Computational Physics*, 231(17):5718–5746, 2012.

[29] B.A. Lockwood and M. Anitescu. Gradient-enhanced universal kriging for uncertainty propagation. *Nuclear Science and Engineering*, 170(2):168–195, 2012.

[30] I. Bilionis, N. Zabaras, B.A. Konomi, and G. Lin. Multi-output separable gaussian process: Towards an efficient, fully bayesian paradigm for uncertainty quantification. *Journal of Computational Physics*, 241:212–239, 2013.

[31] I. Bilionis and I. Zabaras. Solution of inverse problems with limited forward solver evaluations: a bayesian perspective. *Inverse Problems*, 30(1), 2014.

[32] C.E. Rasmussen and C.K.I. Williams. *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press, 2005.

[33] A. O'Hagan. Bayes-hermite quadrature. *Journal of Statistical Planning and Inference*, 29(3):245–260, 1991.

[34] A. O'Hagan, M.C. Kennedy, and J.E. Oakley. Uncertainty analysis and other inference tools for complex computer codes. *Bayesian Statistics 6*, pages 503–524, 1999.

[35] J. Oakley and A. O'Hagan. Bayesian inference for the uncertainty distribution of computer model outputs. *Biometrika*, 89(4):769–784, 2002.

[36] J.E. Oakley and A. O'Hagan. Probabilistic sensitivity analysis of complex models: a bayesian approach. *Journal of the Royal Statistical Society Series B-Statistical Methodology*, 66:751–769, 2004.

[37] P. Chen, N. Zabaras, and I. Bilionis. Uncertainty propagation using infinite mixture of gaussian processes and variational bayesian inference. *Journal of Computational Physics*, 284:291–333, 2015.

[38] I. Bilionis and N. Zabaras. *Bayesian uncertainty propagation using Gaussian processes*. Springer, 2016 (accepted).

[39] D.J.C. Mackay. Bayesian interpolation. *Neural Computation*, 4(3):415–447, 1992.

[40] P. Balaprakash, R.B. Gramacy, and S.M. Wild. Active-learning-based surrogate models for empirical performance tuning, 2013.

[41] D.R. Jones. A taxonomy of global optimization methods based on response surfaces. *Journal of Global Optimization*, 21(4):345–383, 2001.

[42] M. Emmerich, N. Beume, and B. Naujoks. An emo algorithm using the hypervolume measure as selection criterion. *Evolutionary Multi-Criterion Optimization*, 3410:62–76, 2005.

[43] Y. Bengio, O. Delalleau, and N. Le Roux. The curse of highly variable functions for local kernel machines. In *Neural Information Processing Systems*, 2005.

[44] R. Bellman. Dynamic programming and lagrange multipliers. *Proceedings of the National Academy of Sciences of the United States of America*, 42(10):767–769, 1956.

[45] A. Saltelli, M. Ratto, T. Andres, F. Campolongo, J. Caribon, D. Gatelli, M. Saisana, and S. Tarantola. *Global sensitivity analysis: the primer*. John Wiley & Sons, 2008.

[46] R.C. Smith. *Uncertainty quantification: Theory, implementation, and applications*. SIAM, 2014.

[47] R.M. Neal. *Bayesian learning for neural networks.*, volume 118. Springer, New York, 1996.

[48] R.M. Neal. *Assessing relevance determination methods using DELVE*, pages 97–129. Springer-Verlag, 1998.

[49] R. Ghanem and P.D. Spanos. *Stochastic finite elements: a spectral approach*. Dover Publications, Minneola, N.Y., 2003.

[50] K. Person. On lines and planes of closest fit to system of points in space, 1901.

[51] X. Ma and N. Zabaras. Kernel principal component analysis for stochastic input model generation. *Journal of Computational Physics*, 230(19):7311–7331, 2011.

[52] H. Rabitz and O.F. Alis. General foundations of high-dimensional model representations. *Journal of Mathematical Chemistry*, 25(2-3):197–233, 1999.

[53] O.F. Alis and H. Rabitz. Efficient implementation of high dimensional model representations. *Journal of Mathematical Chemistry*, 29(2):127–142, 2001.

[54] G.Y. Li, C. Rosenthal, and H. Rabitz. High dimensional model representations. *Journal of Physical Chemistry A*, 105(33):7765–7777, 2001.

[55] G.Y. Li, S.W. Wang, C. Rosenthal, and H. Rabitz. High dimensional model representations generated from low dimensional data samples. 1. mp-cut-hdmr. *Journal of Mathematical Chemistry*, 30(1):1–30, 2001.

[56] R. Chowdhury, B.N. Rao, and A.M. Prasad. High-dimensional model representation for structural reliability analysis. *Communications in Numerical Methods in Engineering*, 25(4):301–337, 2009.

[57] X. Ma and N. Zabaras. An adaptive high-dimensional stochastic model representation technique for the solution of stochastic partial differential equations. *Journal of Computational Physics*, 229(10):3884–3915, 2010.

[58] J. Wei, G. Lin, L.J. Jiang, and Y. Efendiev. Analysis of variance-based mixed multiscale finite element method and applications in stochastic two-phase flows. *International Journal for Uncertainty Quantification*, 4(6):455–477, 2014.

[59] Z.W. Zhang, X. Hu, T.Y. Hou, G. Lin, and M.K. Yan. An adaptive anova-based data-driven stochastic method for elliptic pdes with random coefficient. *Communications in Computational Physics*, 16(3):571–598, 2014.

[60] T.A. Plate. Accuracy versus interpretability in flexible modeling: Implementing a tradeoff using gaussian process models. *Behaviormetrika*, 26:29–50, 1999.

[61] C.G. Kaufman and S.R. Sain. Bayesian functional ANOVA modeling using Gaussian process prior distributions. *Bayesian Analysis*, 5(1):123–149, 2010.

[62] N. Durrande, D. Ginsbourger, and O. Roustant. Additive covariance kernels for high-dimensional Gaussian process modeling. *arXiv:1111.6233*, 2011.

[63] D. Duvenaud, H. Nickisch, and C.E. Rasmussen. Additive Gaussian processes. In *Advances in Neural Information Processing Systems 24*, pages 226–234.

[64] E. Gliboa, Y. Saatci, and J. Cunningham. Scaling multidimensional inference for structured Gaussian processes. In *30th International Conference on Machine Learning*.

[65] X. Nguyen and A.E. Gelfand. Bayesian nonparametric modeling for functional analysis of variance. *Annals of the Institute of Statistical Mathematics*, 66(3):495–526, 2014.

[66] T.M. Russi. *Uncertainty quantification with experimental data and complex system models*. PhD thesis, 2010.

[67] P. Constantine, E. Dow, and Q. Wang. Active subspace methods in theory and practice: Applications to kriging surfaces. *SIAM Journal on Scientific Computing*, 36(4):A1500–A1524, 2014.

[68] P. Constantine. A quick-and-dirty check for a one-dimensional active subspace, 2014.

[69] P. Constantine, M. Emory, F. Palacios, N. Kseib, and G. Iaccarino. Quantification of margins and uncertainties using an active subspace method for approximating bounds, 2013.

[70] P. Constantine and D. Gleich. Computing active subspaces with monte carlo. *arXiv Pre-print*, 2014.

[71] H. Chen, Q. Wang, R. Hu, and P. Constantine. Conditional sampling and experiment design for quantifying manufacturing error of transonic airfoil. In *Proceedings of the 49th Aerospace Sciences Meeting*, 2011.

[72] P. Constantine, Q. Wang, A. Doostan, and G. Iaccarino. A surrogate accelerated bayesian inverse analysis of the hyshot ii flight data. In *Proceedings of the 49th Aerospace Sciences Meeting*, 2011.

[73] P. Constantine, B. Zaharatos, and M. Campanelli. Discovering an active subspace in a single-diode solar cell model. *arXiv preprint arXiv:1406.7607*, 2014.

[74] T. Lukaczyk, F. Palacios, J. Alonso, and P. Constantine. Active subspaces for shape optimization. In *10th AIAA Multidisciplinary Design Optimization Conference*, 2014.

[75] J. Jefferson, J. Gilbert, P. Constantine, and R. Maxwell. Reprint of: Active subspaces for sensitivity analysis and dimension reduction of an integrated hydrologic model. *Computers & Geosciences*, 2015.

[76] R.E. Plessix. A review of the adjoint-state method for computing the gradient of a functional with geophysical applications. *Geophysical Journal International*, 167(2):495–503, 2006.

[77] A. Griewank and A. Walther. *Evaluating derivatives: principles and techniques of algorithmic differentiation*. Siam, 2008.

[78] P. Geladi and B.R. Kowalski. Partial least-squares regression - a tutorial. *Analytica Chimica Acta*, 185:1–17, 1986.

[79] S. Wold, M. Sjostrom, and L. Eriksson. PLS-regression: a basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems*, 58(2):109–130, 2001.

[80] R.G. Brereton. *Chemometrics: data analysis for the laboratory and chemical plant*. Wiley, 2003.

[81] A.P. Ferreira and M. Tobyn. Multivariate analysis in the pharmaceutical industry: enabling process understanding and improvement in the PAT and QbD era. *Pharmaceutical development and technology*, 20(5):513–27, 2015.

[82] S. Kim., M. Kano, H. Nakagawa, and S. Hasebe. Estimation of active pharmaceutical ingredients content using locally weighted partial least squares and statistical wavelength selection. *International Journal of Pharmaceutics*, 421(2):269–274, 2011.

[83] R.H. Byrd, P.H. Lu, J. Nocedal, and C.Y. Zhu. A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing*, 16(5):1190–1208, 1995.

[84] K. Murphy. *Machine Learning: A Probabilistic Perspective*. Adaptive computation and machine Learning. MIT Press, Cambridge, MA, 2012.

[85] P.W. Goldberg, C.K.I. Williams, and C.M. Bishop. Regression with input-dependent noise: A gaussian process treatment. *Advances in Neural Information Processing Systems 10*, 10:493–499, 1998.

[86] C. Plagemann, K. Kersting, and W. Burgard. Nonstationary gaussian process regression using point estimates of local smoothness. *Machine Learning and Knowledge Discovery in Databases, Part Ii, Proceedings*, 5212:204–219, 2008.

[87] N. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller, and E. Teller. Equation of state calculations by fast computing machines. *Journal of Chemical Physics*, 21(6):1087–1092, 1953.

[88] W.K. Hastings. Monte-carlo sampling methods using markov chains and their applications. *Biometrika*, 57(1):97–109, 1970.

[89] H. Haario, M. Laine, A. Mira, and E. Saksman. Dram: Efficient adaptive mcmc. *Statistics and Computing*, 16(4):339–354, 2006.

[90] T. Lukaczyk, F. Palacios, J. Alonso, and P. Constantine. Active subspaces for shape optimization. In _ *10th AIAA Multidisciplinary Design Optimization Conference* _, 2014.

[91] E. Dow and Q. Wang. Output based dimensionality reduction of geometric variability in compressor blades. In *Proceedings of the 51st Aerospace Sciences Meeting*, 2013.

[92] G.H. Golub and C.F. Van Loan. *Matrix computations*. Johns Hopkins studies in the mathematical sciences. Johns Hopkins University Press, Baltimore, 3rd edition, 1996.

[93] Z. Wen and W. Yin. A feasible method for optimization with orthogonality constraints. *Mathematical Programming*, 142(1-2):397–434, 2013.

[94] R.P. Brent. *Algorithms for minimization without derivatives*. Courier Corporation, 2013.

[95] A. Gelman, J. Hwang, and A. Vehtari. Understanding predictive information criteria for bayesian models. *Statistics and Computing*, 24(6):997–1016, 2014.

[96] The GPy authors. Gpy: A gaussian process framework in python, 2012–2014.

[97] V.F. Nesterenko. *Dynamics of heterogeneous materials*. Springer Science & Business Media, 2013.

[98] V.F. Nesterenko. Propagation of nonlinear compression pulses in granular media. *Journal of Applied Mechanics and Technical Physics*, 24(5):733–743, 1983.

[99] A.N. Lazaridi and V.F. Nesterenko. Observation of a new type of solitary waves in a one-dimensional granular medium. *Journal of Applied Mechanics and Technical Physics*, 26(3):405–408, 1985.

[100] C. Coste, E. Falcon, and S. Fauve. Solitary waves in a chain of beads under hertz contact. *Physical Review E*, 56(5):6104, 1997.

[101] U. Harbola, A. Rosas, M. Esposito, and K. Lindenberg. Pulse propagation in tapered granular chains: An analytic study. *Physical Review E*, 80(3):031303, 2009.

[102] J.J. Lydon. *Nonlinear Effects in Granular Crystals with Broken Periodicity*. PhD thesis, California Institute of Technology, 2015.

[103] M. Girolami and B. Calderhead. Riemann manifold langevin and hamiltonian monte carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(2):123–214, 2011.

[104] P.J. Green. Reversible jump markov chain monte carlo computation and bayesian model determination. *Biometrika*, 82(4):711–732, 1995.

VITA

VITA

Rohit Kaushal Tripathy was born in Sunabeda, Odisha(India) on October 19, 1992. After completing his schooling from DAV Public School in Cuttack, Odisha in 2010, Rohit entered the B. Tech. Mechanical Engineering program at VIT University, Vellore, Tamil Nadu(India). After finishing his undergrad in Vellore, he entered graduate school in the MS Mechanical Engineering program at Purdue University, West Lafayette(USA). Since Fall 2014, Rohit has been working at the Predictive Science Lab in the Mechanical Engineering department at Purdue with Prof. Ilias Bilionis and from Spring 2016, he will begin his doctoral studies in the same research group.